Naval Research Laboratory

Washington, DC 20375-5320

NRL/FR/5550--94-9743

Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform

George S. Kang
Lawrence J. Fransen

Transmission Technology Branch
Information Technology Division

November 9, 1994

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget. Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (*Leave Blank*)

2. REPORT DATE

November 9, 1994

3. REPORT TYPE AND DATES COVERED

Continuing 1 Oct. 1992 - 30 Sep. 1993

4. TITLE AND SUBTITLE

Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform

5. FUNDING NUMBERS

61153N
33904N

6. AUTHOR(S)

George S. Kang and Lawrence J. Fransen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/FR/5550--94-9743

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

This report describes a new speech analysis/synthesis method. This new technique does not attempt to model the human speech production mechanism. Instead, we represent the speech waveform directly in terms of the speech waveform defined in a pitch period. A significant merit of this approach is the complete elimination of pitch interference because each pitch-synchronously segmented waveform does not include a waveform discontinuity. One application of this new speech analysis/synthesis method is the alteration of speech characteristics directly on raw speech. With the increased use of man-made speech in tactical voice message systems and virtual reality environments, such a speech generation tool is highly desirable. Another application is speech encoding operating at low data rates (2400 b/s or less). According to speech intelligibility tests, our new 2400 b/s encoder outperforms the current 2400-b/s LPC. This is also true in noisy environments. Because most tactical platforms are noisy (e.g., helicopter, high-performance aircraft, tank, destroyer), our 2400-b/s speech encoding technique will make tactical voice communication more effective; it will become an indispensable capability for future C41.

14. SUBJECT TERMS

Speech modeling	Speech synthesis
Speech alterations	Pitch-synchronous segmentation of speech waveform
Speech coding	

15. NUMBER OF PAGES

55

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

SPEECH ANALYSIS AND SYNTHESIS BASED ON PITCH-SYNCHRONOUS SEGMENTATION OF THE SPEECH WAVEFORM

INTRODUCTION

The Operation Desert Storm experience led the Joint Chiefs of Staff to create the *C4I for the Warrior* concept, which is a road map for achieving future communication and intelligence systems. As General Colin Powell, former chairman of Joint Chiefs of Staff (JCS) stated, the *C4I for the Warrior* concept will give the battlefield commander access to all information needed to win the war and will provide the information when, where, and how the commander wants it [1]. Several DoD directives support the *C4I for the Warrior* concept. We list three examples:

1. **JCS's *Top Five Future Joint Warfighting Capabilities*** [2]: Significantly, the first wanted capability listed in this document is directly related to communication technology:

   > To maintain near-perfect real-time knowledge of the enemy and communicate that to all forces in near real-time.

2. **ONR's *Naval Needs and Scientific Opportunities*** [3]: The first item listed in this document is again communication:

   > Global surveillance and communication.

3. **Navy's *Copernicus* Architecture** [4]: Similarly, the technology needed to implement *Copernicus* (the Navy C4I goal architecture) includes items related to voice communication:

   > Low-data-rate voice below 2400 bits/second.

Voice communication technology, particularly low-data-rate voice communication technology, is vital to future C4I because tactical commanders still rely on narrowband links. Currently, low-data-rate voice communication is supported by the Advanced Narrowband Voice Terminal (ANDVT) and the Secure Voice Terminal (STU-III). Both terminals use the linear predictive coder (LPC) to encode speech at 2400 b/s. Although LPC provides better speech intelligibility and quality than previously deployed voice coding systems, at least three limitations are associated with the LPC:

- *Lower intelligibility for female voices*: Speech intelligibility of the female voice encoded by the 2400-b/s LPC is 5.1 points below that of the male voice in various operating conditions. [All intelligibility scores cited in this report were measured with the Diagnostic Rhyme Test

Manuscript approved August 22, 1994.

(DRT).] Because of the increasing number of females engaged in combat operations, this limitation cannot be overlooked.

- *Poor speaker recognition*: A familiar voice encoded by the 2400-b/s LPC is not easily recognized. According to testing conducted by Astrid Schmidt-Nielsen of NRL, the 2400-b/s LPC speaker recognition score is only 69% [5]. Speaker recognizability is important because people tend to be reluctant to talk to unidentified communicators.

- *Significant loss of speech intelligibility with noisy speech*: Intelligibility of LPC-processed speech quickly degrades if the speech is contaminated by acoustic noise. For example, intelligibility of LPC-processed speech that originates at a helicopter platform is only in the 60% range, which is about 15 points below the minimum acceptable level of speech intelligibility. Since military platforms (e.g., high-performance aircraft, helicopter, tank, destroyer, armored personnel carriers) are generally noisy, the effectiveness of tactical voice communication is hindered by the limitations of current narrowband voice terminals.

In an attempt to improve the effectiveness of low-data-rate voice coding, we have developed a new speech analysis/synthesis technique. This report describes this new voice analysis/synthesis technique with two major applications: (1) voice alterations for man-made speech, and (2) low-data-rate voice encoding for tactical voice communication. This report can be summarized as follows:

- *New speech model*: Our new speech model is fundamentally different from the speech model that has been used for the past 50 years. The existing speech model is an electric analog of the human voice production mechanism and consists of a filter and an excitation signal. In the old model, the parameters associated with the filter and excitation represent the speech waveform. As known from the solution to system characterization problems, however, the estimation of system parameters requires both the input to and output from the system. The problem with the old model is that we have no access to the system input (turbulent air from the lungs). This results in an inaccurate estimation of the system parameters. To avoid this difficulty, we developed a completely new speech modeling approach in which the speech waveform is directly represented.

- *Capability to alter raw speech*: At present, speech alterations are possible only when using synthetic speech generated by the existing speech model. In contrast, our new speech model permits alterations of speech characteristics (utterance rate, pitch, and resonant frequencies) using raw speech. As a result, our altered speech sounds more natural. The capability to alter speech characteristics is becoming an important speech processing tool with the increased use of man-made speech in tactical communication and virtual-reality environments. Our new speech analysis/synthesis method will play a significant role in this application.

- *Improved speech coding*: The speech model plays a vital role in low-data-rate speech coding because speech is transmitted in terms of speech parameters and then synthesized at the receiver. As stated earlier, speech parameters derived from the old speech model are not accurate. This is particularly true for high-pitched female voices as the result of the high rate of pitch interference. Our new speech model is not affected by pitch interference because the speech waveform is pitch-synchronously segmented prior to analysis. Thus, the encoded speech sounds better at a comparable data rate. This report shows that female speech intelligibility of our new approach is superior to LPC at 2400 b/s. Thus, the low-data-rate voice encoder based on our new speech model is ideal for narrowband tactical communications.

The new speech analysis technique described in this report can be exploited in the current LPC for improved speech intelligibility. This approach is described in the appendix.

This report is a product of our continued effort to improve tactical voice communication. It was written for three groups of people: program managers and sponsors who are actively involved in the transfer of voice technology to working hardware; communication-architecture planners who are interested in the state-of-the-art of voice transmission; and independent researchers who develop voice terminals.

BACKGROUND

This section compares the traditional speech model and our new approach to show that there are fundamental differences between the two approaches.

Existing Speech Model

Homer Dudley developed a speech model in 1939 [6] that has been extensively used for speech encoding (e.g., ANDVT and STU-III) and speech synthesis (e.g., text-to-speech converters). A goal of any speech model is to convert the complex speech waveform into a set of perceptually significant parameters (e.g., resonant frequencies, pitch, loudness, voicing). In turn, by controlling these speech parameters, speech can be generated by the speech model. Although numerous speech models have been developed in the past, they all basically consist of a filter (representing the vocal tract) driven by an excitation signal (representing the signal from the glottis) (Fig. 1).

In this conventional speech model, model parameters (i.e., vocal tract filter and excitation signal parameters) are derived by the use of the output signal (i.e., the speech signal) alone. As a result, estimated model parameters are often not accurate.

New Speech Model

In our approach, we do not try to generate an electric analog of the human speech production mechanism. Rather, we represent the speech waveform by individual pitch cycle waveforms. The pitch synchronously segmented speech waveforms are structurally similar to the individual frames of a motion picture, which are snapshots of a moving object (Fig. 2(a)). Likewise, each pitch period of the speech waveform is a snapshot of continuous speech (Fig. 2(b)).

Figure 2(b) shows that the speech waveform is discontinuous at the beginning of each pitch cycle because each pitch waveform is generated by a different excitation. However, LPC performs a running autoregressive analysis of the speech waveform to extract speech spectral parameters. This analysis is equivalent to averaging a horse's head and legs to derive a physically meaningful quantity. As expected, LPC results are not that good, particularly for high-pitched female voices where waveform discontinuities occur more frequently.

The new speech model (or speech analysis and synthesis technique) is based on a completely new perspective, based on the structure of a motion picture. We incorporate the desired features directly on the given speech waveform by pitch-synchronous segmentation, time-domain stretching (or shrinking), and frequency-domain stretching (or shrinking). The output speech sounds more natural because it is constructed from raw speech. We also quantized pitch-synchronously segmented waveforms to implement a speech encoder. Output speech sounds more natural because pitch interference is absent in the speech representation.

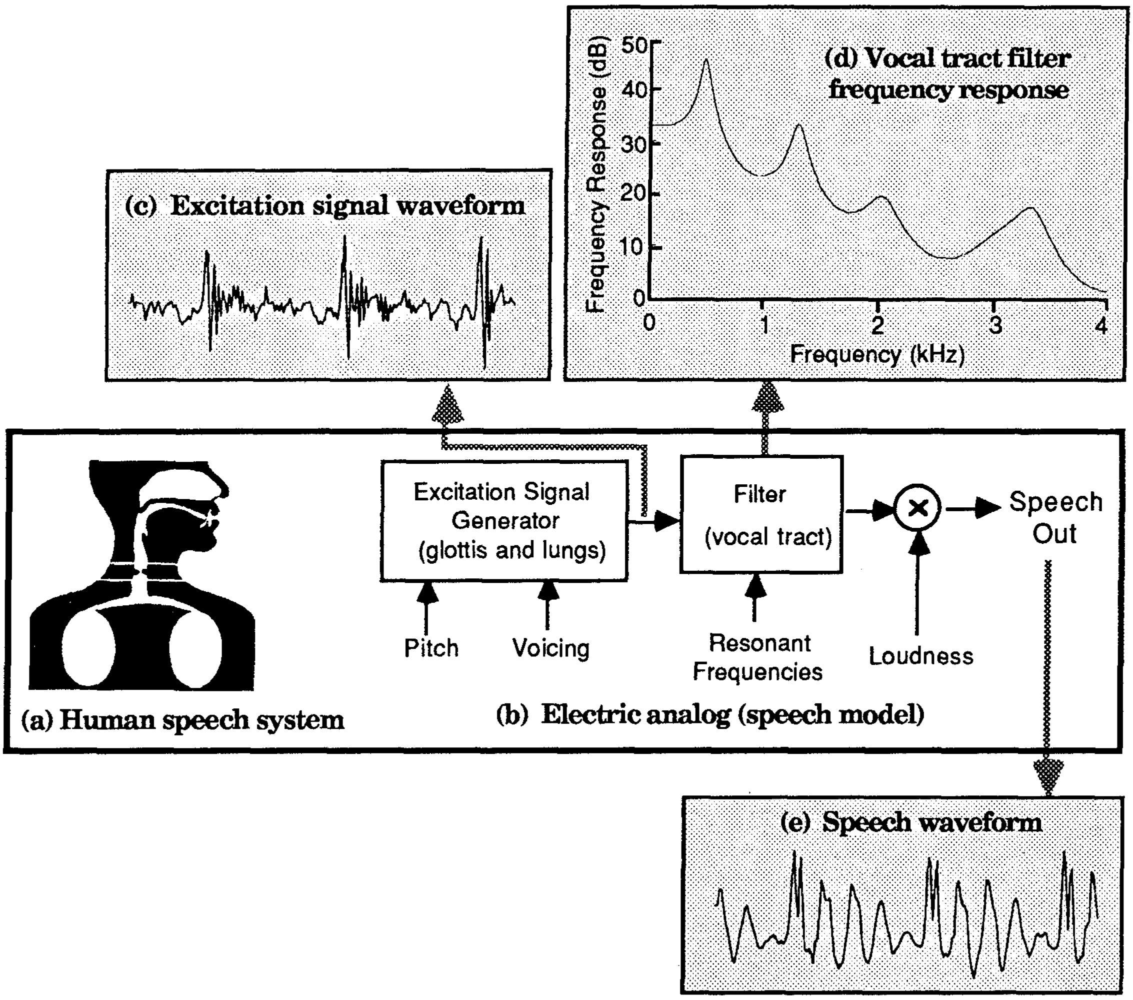
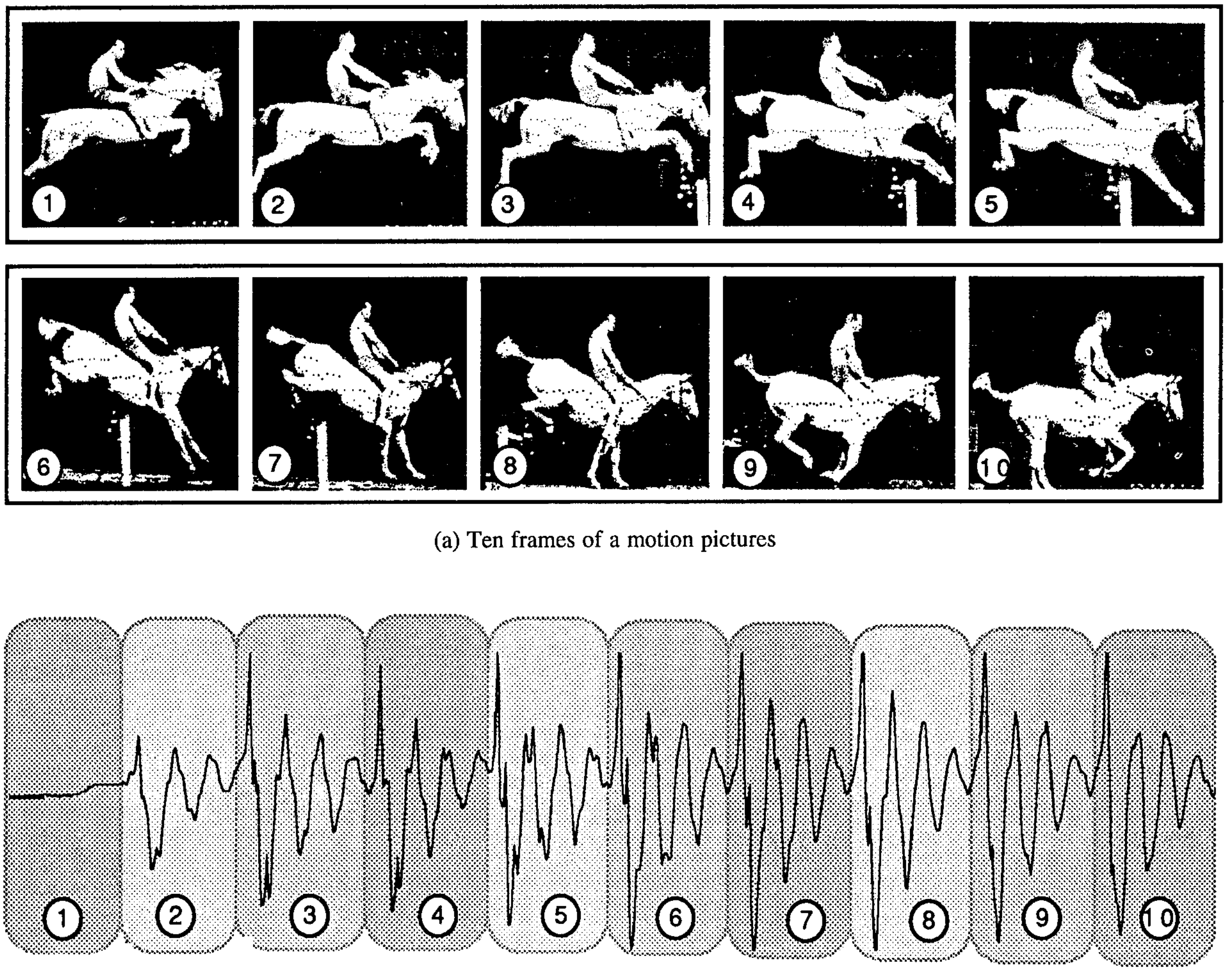


Fig. 1 — Generic speech model based on the human speech production mechanism. The speech waveform (Fig. 1(e)) is the only quantity that is available. The vocal tract filter response (Fig. 1(d)) must be estimated from the speech waveform. The excitation signal (Fig. 1(c)) is modeled by either a pulse train (to generate vowels) or random noise (to generate consonants). The use of such a simple signal enables us to control the pitch period, utterance rate, and resonant frequencies independently.



(a) Ten frames of a motion pictures

(b) Ten pitch cycles of a speech waveform

Fig. 2 — Analogy between moving picture and speech waveform. A moving picture and speech waveform are strikingly similar in structure. Our new technique begins with pitch-synchronous segmentation of the speech waveform.

NEW SPEECH ANALYSIS AND SYNTHESIS METHOD

This section describes the new approach to speech modeling. This new technique does not attempt to model the human speech production mechanism as done previously. We represent the speech waveform directly in terms of the speech waveform defined in a pitch period. A significant merit of the new approach is the complete elimination of pitch interference because each segmented waveform does not include a waveform discontinuity. Figure 3 is a block diagram of the new speech analysis/synthesis technique.

Analysis (Pitch-Synchronous Speech Segmentation)

The purpose of analysis is to synchronously segment the given speech waveform pitch. The speech segmentation must be performed in such a manner that when the segmented waveforms are concatenated, no clicks, rattles, or warbles are generated. The resultant speech sounds must be similar to the original speech, even if only one out of many segmented waveforms are directly used to generate the speech. The speech segmentation technique is the key to the new approach. We describe each block in Fig. 4.

Low-Pass Filter

The low-pass filtering (0 to 1 kHz) removes higher-frequency speech content, which is detrimental to pitch tracking. Figure 5 shows that the periodic vowel waveform has few periodic components in the frequency region above 1 kHz. Removing these random components improves pitch tracking. We implemented a 19-tap, linear-phase, low-pass filter that has a frequency response of –3 dB at 1 kHz with a –60 dB/octave roll-off characteristic. The passband characteristic of the low-pass filter is not as critical as the characteristics of the other blocks in Fig. 4.

Average Magnitude Difference Function

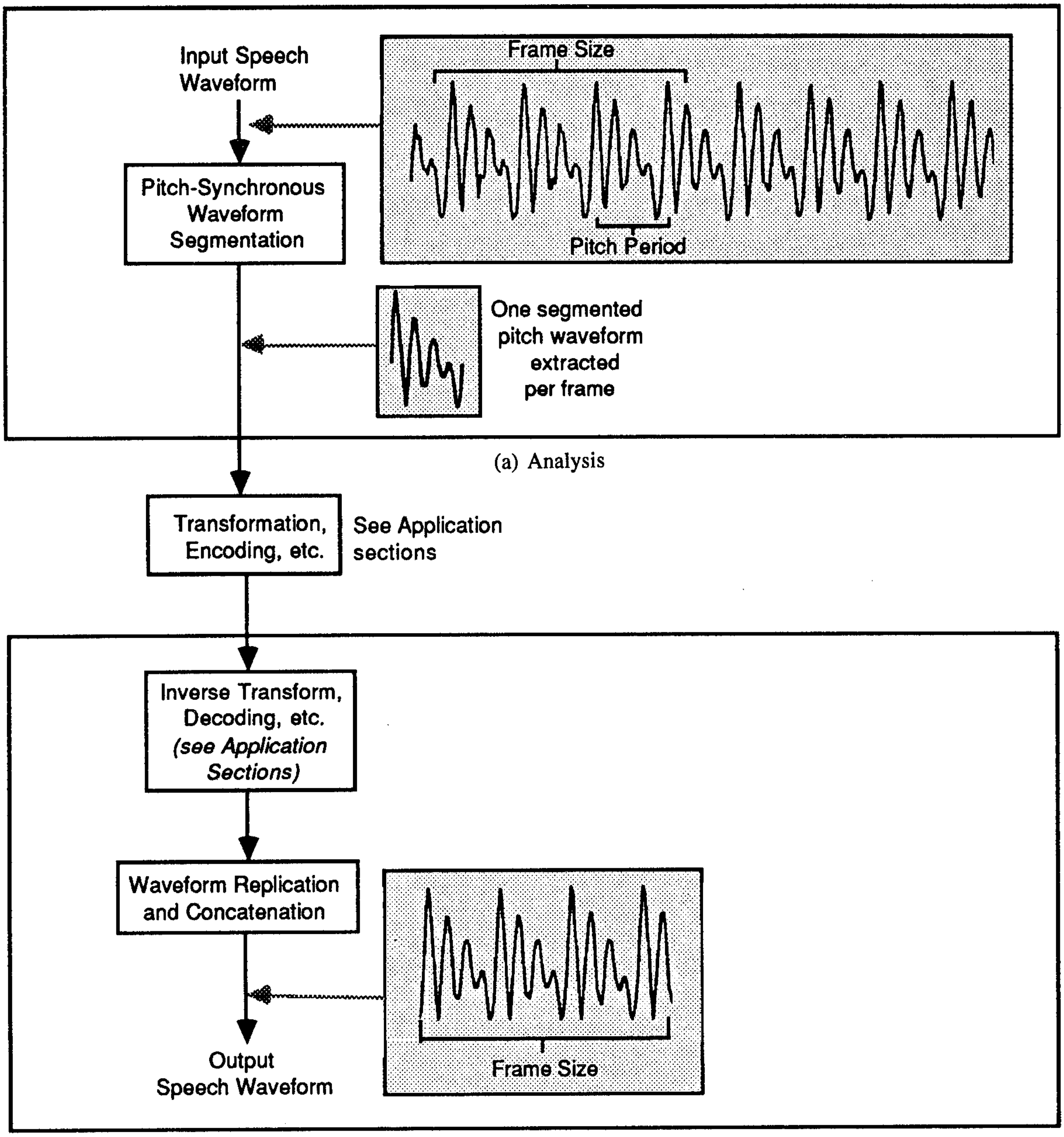
The Average Magnitude Difference Function (AMDF) is a simpler alternative to an autocorrelation function (ACF) that indicates the degree of correlation of the input signal at various mutual separations. AMDF is preferred over ACF from a computational point of view. The dynamic range is on the same order of magnitude as the speech waveform (a sum of differences, rather than a sum of products). The AMDF may be computed by a block-form analysis with a fixed analysis window and an integration and dump operation:

$$A(K,\tau)=\sum_{k=1}^{K} |e(k)-e(k+\tau)|\,, \tag{1}$$

or flow-form analysis with a moving analysis window and recursive updating (low-pass filtering):

$$A(k,\tau)=(1-\beta)\,A(k-1,\tau)+\beta\,|\,e(k)-e(k+\tau)\,|\,, \tag{2}$$

where $e(k)$ is the kth sample of the 1 kHz low-passed speech sample, K is the fixed analysis window size (which is usually equal to frame size), τ is a mutual time separation, and β is a feedback constant that controls the 3 dB cut-off frequency. If β = .01, the 3 dB corner frequency of 45 Hz is on the same order as the frame rate. Although the number of computational steps is greater, the flow-form analysis provides a smoother AMDF histogram, hence more reliable pitch tracking.


Input Speech Waveform
Frame Size
Pitch-Synchronous Waveform Segmentation
Pitch Period
One segmented pitch waveform extracted per frame
Transformation, Encoding, etc.
See Application sections
Inverse Transform, Decoding, etc. *(see Application Sections)*
Waveform Replication and Concatenation
Frame Size
Output Speech Waveform

(a) Analysis

(b) Synthesis

Fig. 3 — The new speech analysis/synthesis technique; speech waveform segmentation and concatenation are clearly shown. When the frame size is larger than a pitch period, the segmented speech waveform is replicated to regenerate the output speech. This segmented speech waveform can be replicated two or three times without degrading speech intelligibility.

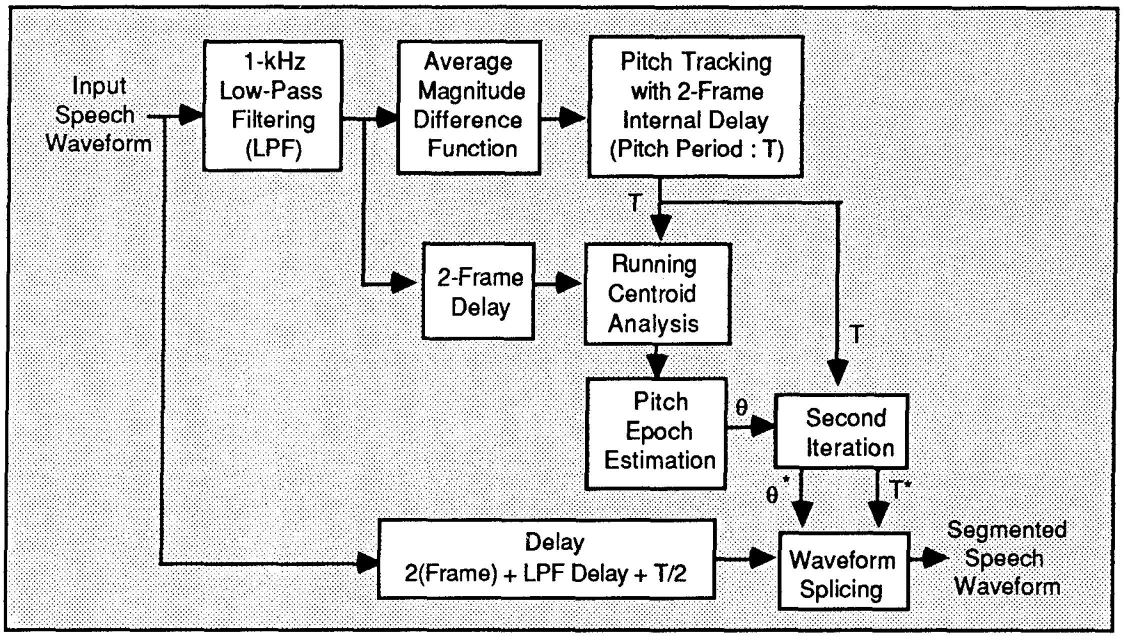

Input Speech Waveform
1-kHz Low-Pass Filtering (LPF)
Average Magnitude Difference Function
Pitch Tracking with 2-Frame Internal Delay (Pitch Period : T)
T
2-Frame Delay
Running Centroid Analysis
T
Pitch Epoch Estimation
θ
Second Iteration
θ*
T*
Delay 2(Frame) + LPF Delay + T/2
Waveform Splicing
Segmented Speech Waveform

Fig. 4 — Pitch-synchronous speech segmentation (analysis)

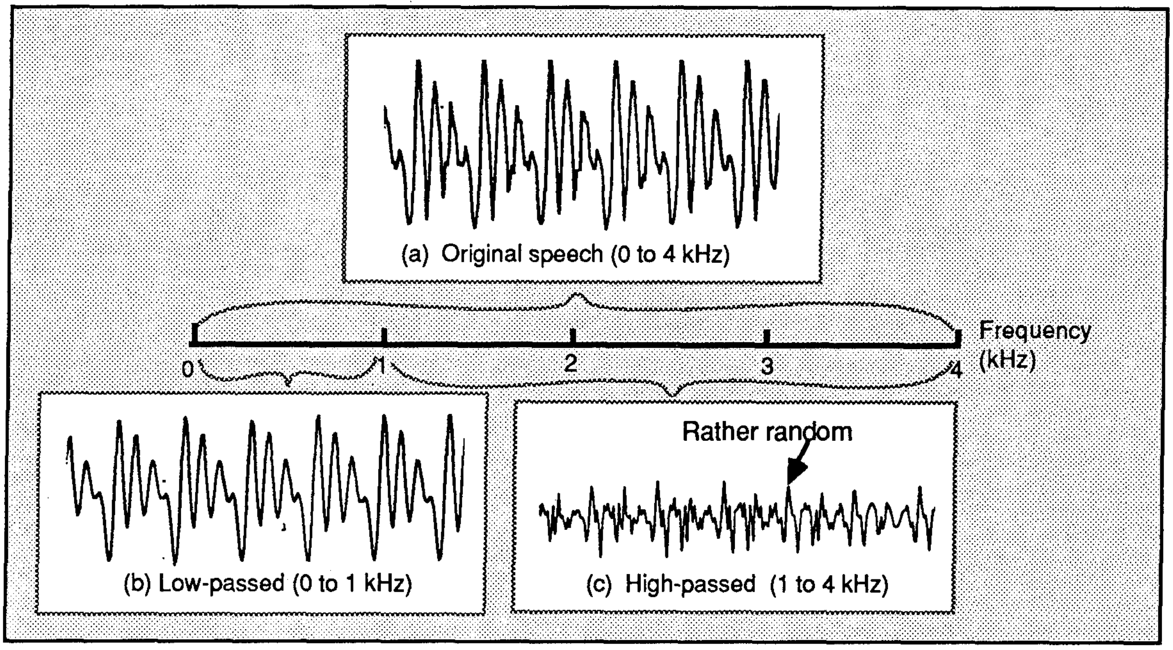
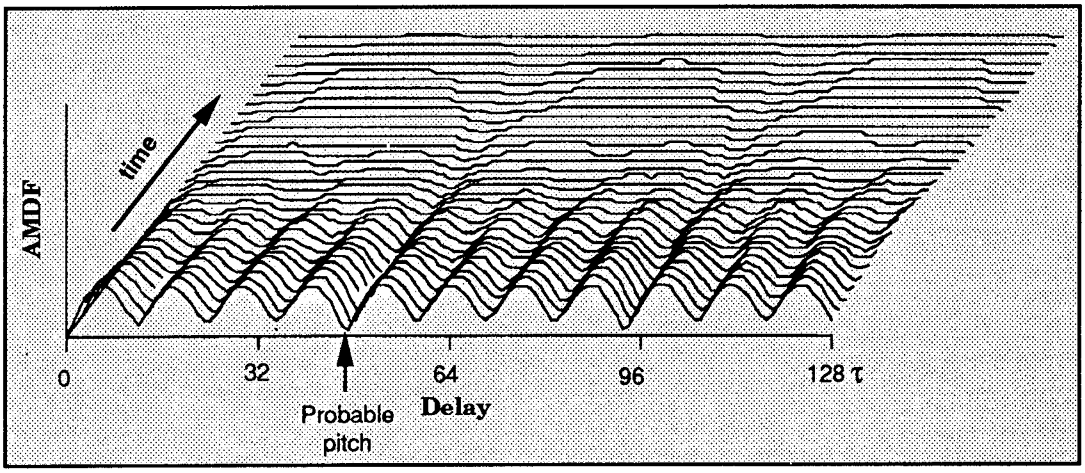


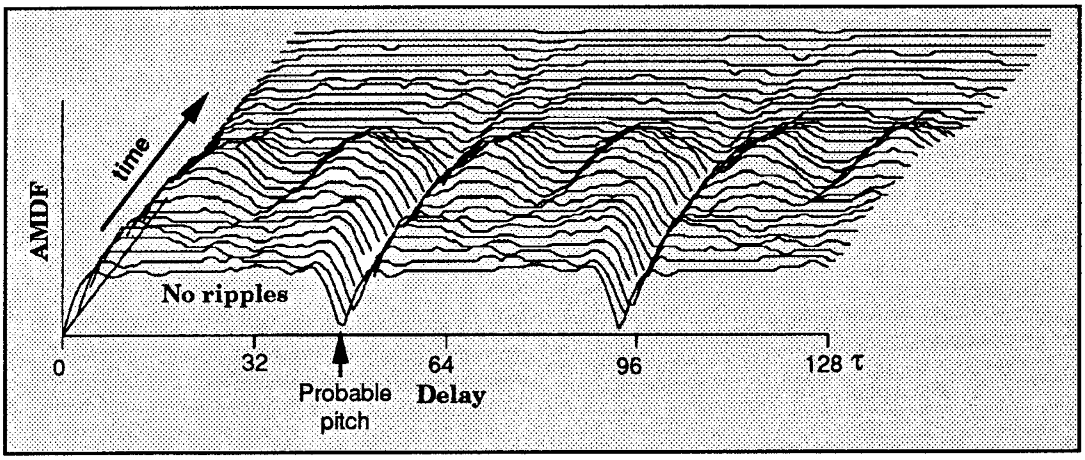
Fig. 5 — Steady vowel in three different frequency bands. Note that high-frequency components of speech are significantly more random. The pitch can be tracked more reliably if these high-frequency components are eliminated.

Figure 6 shows typical AMDFs computed by Eq. (2). The delay corresponds to the minimum AMDF, which is the most probable pitch period, and the magnitude of minimum value is a measure of confidence. However, we do not decide the pitch value directly from the single location of the AMDF minimum; rather, the pitch tracker decides the pitch value based on the history of recent AMDF minima.

Because voiced speech has at least one resonant frequency below 1 kHz, the AMDF will have ripples (formant interference) for voiced speech input, as indicated by Fig. 6(a). If the speech signal is spectrally flattened prior to input to the 1 kHz low-pass filtering and AMDF operations, the formant interference is significantly reduced, as noted in Fig. 6(b). A pitch tracker will perform better if the formant interference is reduced.



(a) Without spectral flattening



(b) With 10-tap spectral flattening

Fig. 6 — AMDFs of a 1 kHz low-passed speech signal with and without spectral flattening. Each trace of AMDF is separated by 40 sampling time periods (5 ms). The delay corresponding to a local minimum is a probable pitch period. As noted, spectral flattening reduces ripples (formant interference) in the AMDF. The AMDF profile is passed to the pitch tracker to estimate pitch trajectory.

Pitch Tracker

The purpose of a pitch tracker is to estimate a continuous pitch trajectory. Any reliable pitch tracker is acceptable for our application. We used the one that is used in the DoD 2400-b/s LPC [7].

Pitch-Epoch Determination by Centroid Analysis

Centroid analysis is used to determine the center of gravity of each pitch waveform. From the centroid, we determine the two adjacent pitch epochs. Our attempts to determine pitch epochs based on an instantaneous value (e.g., the zero crossing before a local peak in the speech waveform) were not successful. We must use all of the speech samples within the analysis window (i.e., one pitch period) to determine pitch epochs, as discussed below.

The center of gravity of a flat object having a mass distribution $f(x)$ (where $f(x) > 0$) is defined as

$$\eta = \frac{\int_{x1}^{x2} xf(x)dx}{\int_{x1}^{x2} f(x)dx} \qquad \text{for } x1 \le x \le x2, \tag{3}$$

which is the ratio of the first moment to the zeroth moment. This concept of the center of gravity has been used in the field of signal analysis in recent years [8].

Based on Eq. (3), we perform the running centroid analysis with an analysis window size equal to the pitch period. For convenience, the time origin is set at the center of the analysis window. Thus, the quantity to be evaluated at each speech sampling time instant is

$$\eta(t) = \frac{\sum_{\tau=t-T/2}^{t+T/2-1} (\tau - t)\,[e(\tau) + e_b]\,d\tau}{\sum_{\tau=t-T/2}^{t+T/2-1} [e(\tau) + e_b]\,d\tau}, \tag{4}$$

where $e(\tau)$ is the 1 kHz low-pass filtered speech signal, and T is the current pitch period obtained from the pitch tracker. Since the speech waveform $e(\tau)$ has both positive and negative values, we must add a bias e_b = 32,767 (for a 16-bit representation of speech samples) so that the quantity inside the bracket is a positive value. An example of a running centroid is plotted in Fig. 7(b), which was generated from the speech waveform shown in Fig. 7(a).

Equation (4) can be simplified for the speech signal in which the DC component (the average value) is removed at the front end. In the absence of a DC component,

$$\sum_{\tau=t-T/2}^{t+T/2-1} e(\tau)\,d\tau = 0. \tag{5}$$

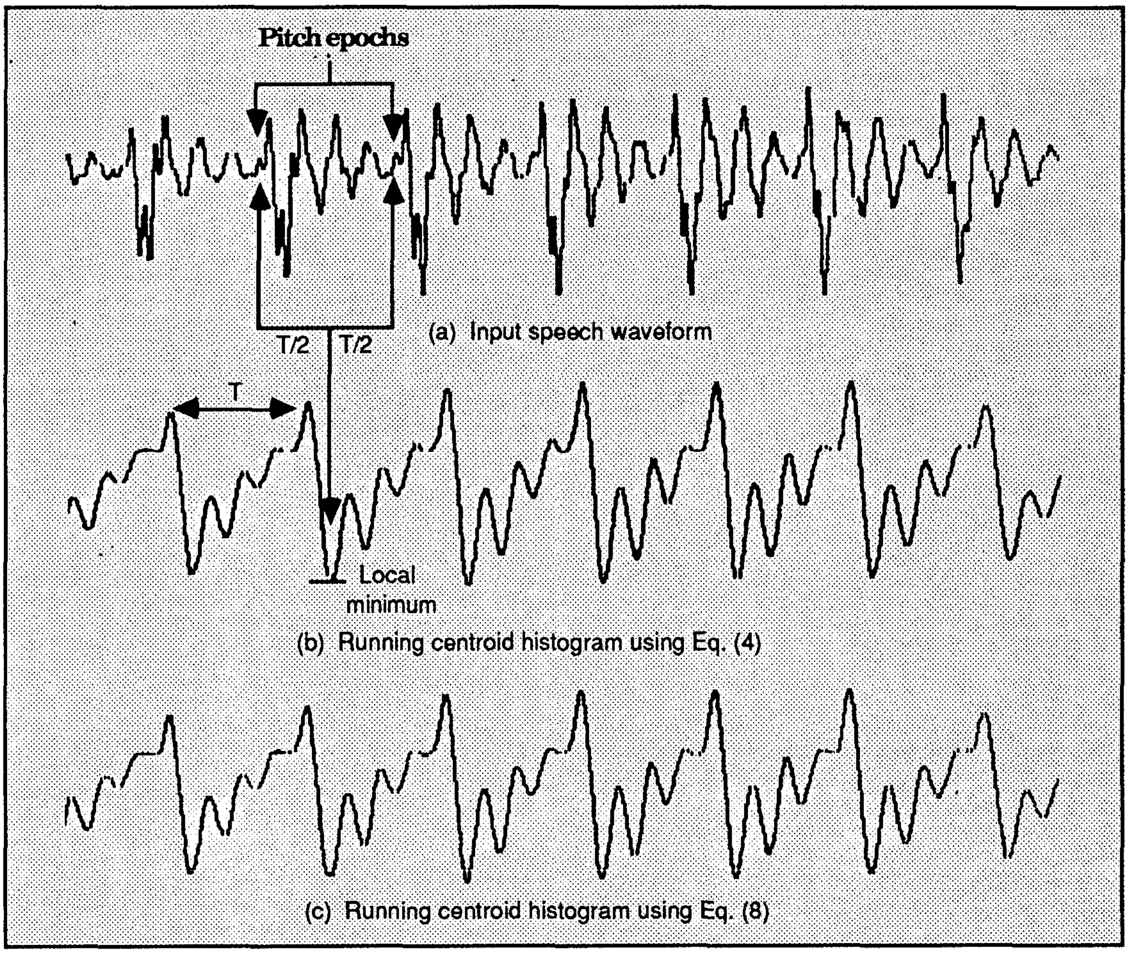


Fig. 7 — Running centroid histogram obtained from one pitch period of the speech signal. Figure 7(b) and Fig. 7(c) are based on Eqs. (4) and (8), respectively. Both results are identical as far as the determination of pitch epochs, and Eq. (8) is much simpler. Pitch epochs are a half pitch period away from the local minimum centroid location.

Furthermore,

$$\sum_{\tau=t-T/2}^{t+T/2-1} e_b \, d\tau = 0 \tag{6}$$

and

$$\sum_{\tau=t-T/2}^{t+T/2-1} (\tau - t)\, e_b \, d\tau = 0 . \tag{7}$$

Thus, Eq. (4) reduces to

$$\eta(t) = G \sum_{\tau=t-T/2}^{t+T/2-1} (\tau - t)\, e(\tau) d\tau , \tag{8}$$

where $G = e_bT$ is a constant for a given frame. An example of a running centroid based on Eq. (8) is plotted in Fig. 7(c), which is essentially the same as Fig. 7(b) stemming from Eq. (4). Since Eq. (8) is computationally simpler than Eq. (4), we use Eq. (8) for subsequent computations.

As noted from Fig. 7(b) or 7(c), the local minimum of the centroid histogram repeats at the rate of the fundamental pitch frequency. The time instant that corresponds to a local minimum is the best centroid for that pitch cycle. The pitch epochs are marked a half pitch period before and after this time location.

Second Iteration

Once the initial pitch epoch estimation is made, a second iteration must be made to account for the small pitch-to-pitch speech waveform variations. We recompute the running centroid histogram by perturbing both the tracked pitch period (T) and the pitch epoch (lower limit of Eq. (11)) by a small amount (±5% of the first estimate). During this iteration, we store the location for the minimum centroid. This second iteration improves the pitch-synchronous segmentation, which significantly improves the quality of the synthesized speech.

Synthesis (Segmented-Speech Replication)

Speech synthesis is accomplished by concatenating pitch-synchronously segmented waveforms. The segmented waveform given by the analyzer is replicated pitch-synchronously. Because the frame size (L) is generally greater than the pitch period (T), the segmented speech waveform is usually replicated more than once (Fig. 8). The segmented waveform is always replicated in its entirety.

Waveform Replication Near the Frame Boundary

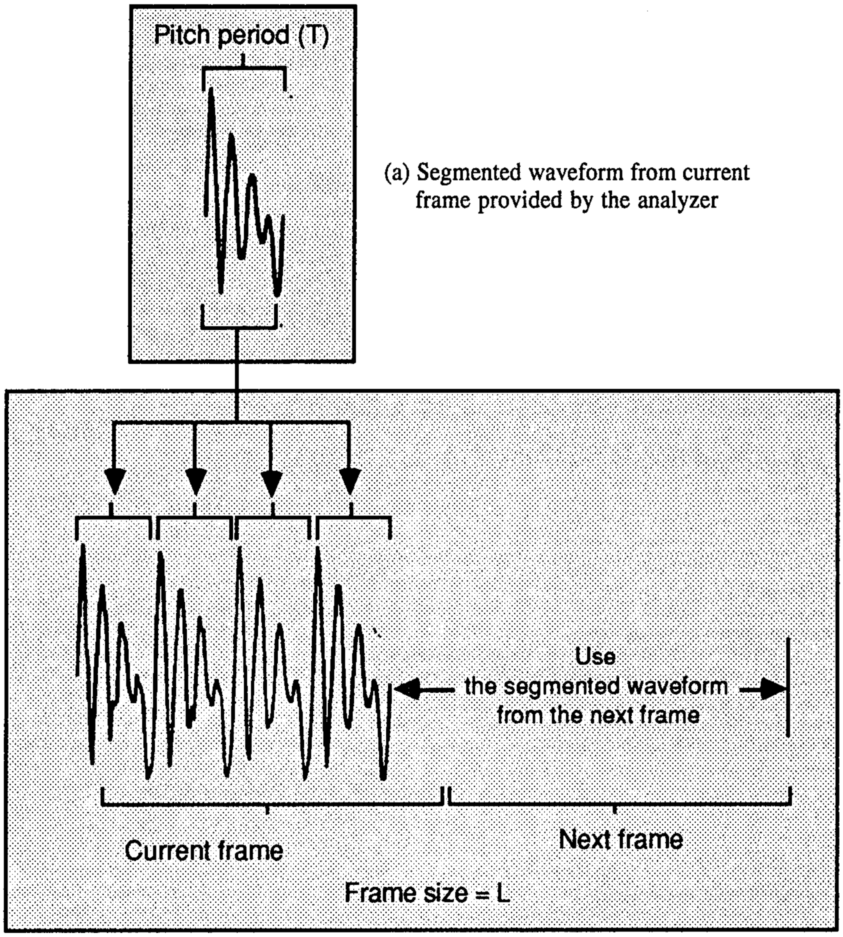
Near the boundary, it is necessary to decide whether the segmented waveform of the current frame should be replicated again or the segmented waveform of the next frame should be copied. The choice is determined by the remaining space in relation to the length of the segmented waveform (T). If the remaining space is greater than $T/2$, the segmented waveform of the current frame is replicated again. On the other hand, if the remaining space is less than or equal to $T/2$, the segmented waveform of the next frame is copied.

End Point Interpolation

Any significant discontinuity at the segmented waveform boundary will produce clicks or warbles in the synthesized waveform. To avoid discontinuities, we perform a mild three-point interpolation at the pitch epoch:

$$e_0(j) = \begin{cases} .25e_i(j-1) + .5e_i(j) + .25e_i(j+1) & \text{if } j = 1 \\ e_i(j) & \text{if } j = 2,3,4,\ldots,T \end{cases}, \tag{9}$$

where $e_i(j)$ is the jth segmented speech sample of the current frame.



(a) Segmented waveform from current frame provided by the analyzer

(b) Concatenated waveform generated by synthesizer

Fig. 8 — Concatenation of segmented speech waveform

Evaluation of Concatenated Speech

Figure 9 compares the given speech waveform and the analyzed-and-synthesized speech waveform. The waveforms are closely matched, requiring careful visual inspection to discern differences between the two.

The Diagnostic Rhyme Test (DRT) was used to evaluate the intelligibility of both the raw speech and the segmented and concatenated speech. The DRT evaluates the discriminability of initial consonants of monosyllable rhyming word pairs. For many years, DRT scores have been widely used as a diagnostic tool to refine voice processors. Likewise, it has been effectively used to rank several competing voice processors. An extensive amount of DRT data has been collected from different voice processors under varied operating conditions. Our experience shows that DRT scores are dependable; scores are repeatable under retesting. They often reveal latent defects in synthetic speech that are not easily discernible through casual listening.

Table 1 lists the DRT scores obtained from the raw speech (0-4 kHz) and the analyzed/synthesized speech. The average DRT scores for 3 male and 3 females speaker are 95.0 for raw speech and 94.7 for analyzed/synthesized speech. The closeness of these two sets of scores are

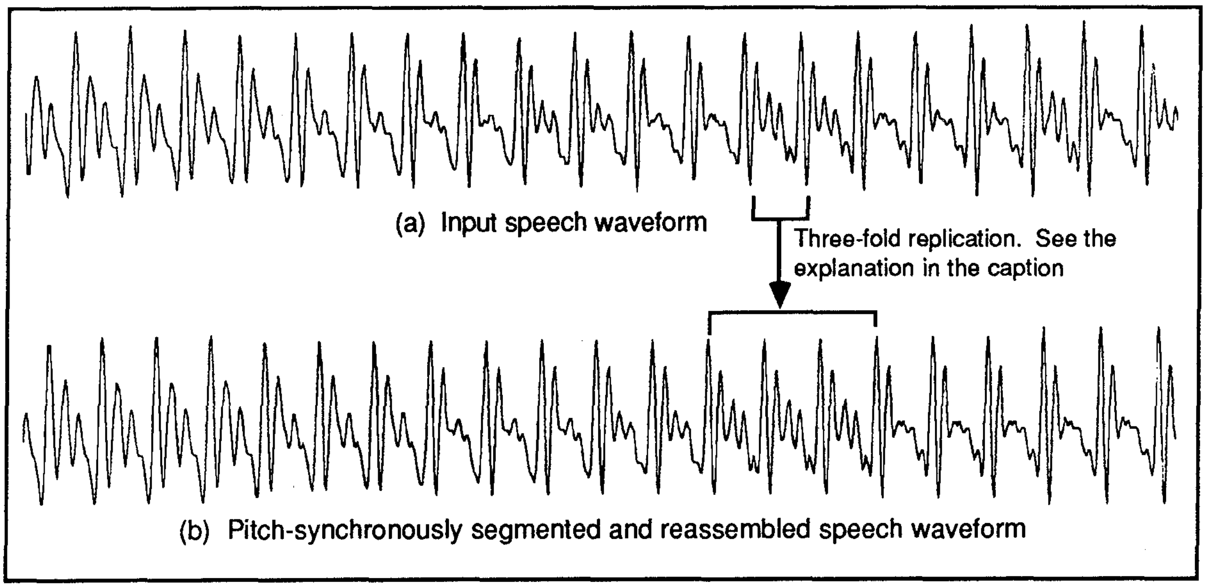

(a) Input speech waveform
Three-fold replication. See the explanation in the caption
(b) Pitch-synchronously segmented and reassembled speech waveform

Fig. 9 — Input and output speech waveforms. Comparison of these two waveforms shows that pitch-synchronous segmentation and concatenation produces a nearly flawless result. Output speech sounds almost as good as input speech. This is an example from a high-pitched voice, pitch period is 44 and frame size is 120. Thus, each segmented speech waveform is repeated approximately three times per frame, on the average.

Table 1 — DRT Scores of Raw Speech and Analyzed/Synthesized Speech

DRT Attribute		DRT Scores			
		3 Male Speakers		3 Female Speakers	
		Raw Speech*	Analyzed & Synthesized**	Raw Speech*	Analyzed & Synthesized***
Voicing	Distinguishes /b/ from /p/, /d/ from /t/, /v/ from /f/, etc.	96.9	96.6	95.1	95.1
Nasality	Distinguishes /n/ from /d/, /m/ from /b/, etc.	96.1	99.2	96.9	99.2
Sustention	Distinguishes /f/ from /p/, /b/ from /v/, /t/ from /θ /, etc.	86.7	91.4	82.8	92.7
Sibilation	Distinguishes /s/ from /θ /, /ʃ / from /d/, etc.	96.4	97.7	95.6	90.9
Graveness	Distinguishes /p/ from /t/, /b/ from /d/, /m/ from /n/, etc.	81.5	87.0	79.9	89.1
Compactness	Distinguishes /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc.	95.1	98.4	97.1	98.7
	TOTAL	96.5	95.1	93.5	94.3

* Bandlimited from 0 to 4 kHz
** Frame = 120 samples
*** Frame = 80 samples

Compare Compare

remarkable considering that a significant portion of the speech waveform was eliminated in the generation of the analyzed/synthesized speech.

The average pitch period for male and female speakers is 72 and 36, respectively; the analysis frame size is 120 and 80, respectively. Thus, the average number of waveform replications is 1.67 (=120/72) for male voices and 2.22 (=80/36) for female voices. Although replication reduces the information in output speech, speech intelligibility is virtually unimpaired with the new analysis and synthesis technique. In other words, the analyzed-and-synthesized speech waveform is almost as intelligible as raw speech itself.

APPLICATION TO VOICE MODIFICATION

Once the speech waveform has been segmented, the individual pieces are usually transformed or encoded, depending on the application. Transformations modify the utterance rate, pitch period, or resonant frequencies of the speech to produce a special effect. Encoding involves quantizing each segmented waveform for use in low-data-rate voice communication systems. This section describes methods of speech transformation; the next section discusses encoding.

One important application of a speech model is for altering the characteristics of stored speech samples for speech generation. This would be extremely useful in such contexts as tactical message systems that transmit text and generate speech at the receiver. A majority of one-way tactical messages can be transmitted in this manner at extremely low data rates (even below 100 b/s) [9]. Because the required data rate is so low, this approach is well-suited for implementing an acoustic "telephone" between submarines, or between submarines and surface ships.

Existing speech analysis/synthesis techniques can only modify the characteristics of speech after it has been synthesized. This results in speech that lacks sufficient quality and intelligibility. However, the new speech model described here allows us to directly modify the characteristics of the raw speech. The difference is very significant. We discuss three speech modification techniques:

- utterance rate modification,
- pitch period modification, and
- resonant frequency modification.

Utterance Rate Modification

If recorded speech is played back faster, both the pitch and the resonant frequencies are increased proportionally. The resultant speech sounds like Donald Duck. When recorded speech is played back at a slower rate, the reverse is true: the resultant speech becomes slurred. When speech is generated by a conventional speech model, however, the speech rate is independent of pitch or resonant frequencies. This is also true with our new speech analysis approach. The difference is that we incorporate speech modification directly on the given speech rather than on synthetic speech; therefore, our speech sounds more intelligible. The speech utterance rate can be altered by simply altering the synthesis frame size relative to the analysis frame size:

$$\text{Output Speech Rate} = (L1/L2)\,(\text{Input Speech Rate}), \qquad (10)$$

where $L1$ and $L2$ are the input and output frame sizes, respectively (Fig. 10). The frame size is one of the constants of the speech analysis/synthesis software.

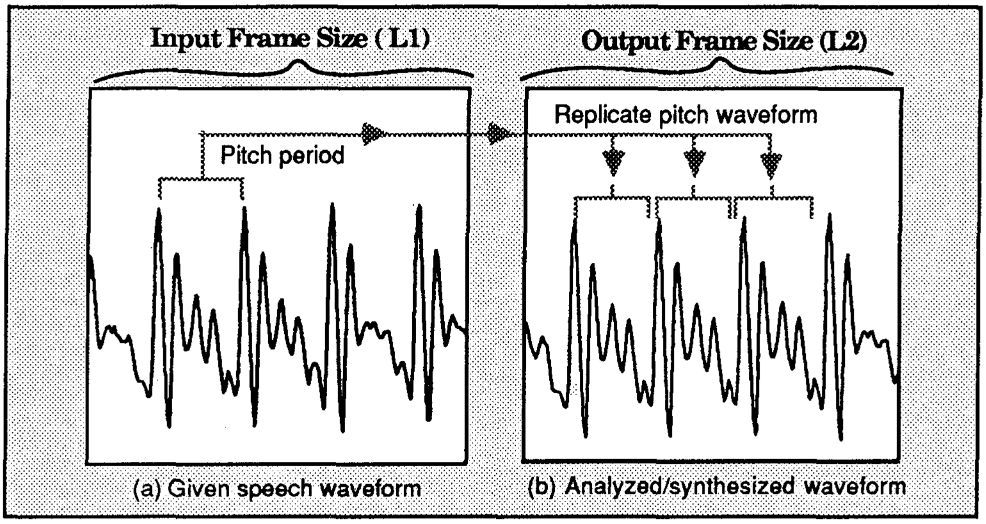


Fig. 10 — Input and output frames. The ratio of the input frame size ($L1$) to output frame size ($L2$) determines the output speech rate. When $L2 = L1$, the output speech rate equals the input speech rate. When $L1/L2 < 1$, the output speech rate is reduced by the same factor. When $L1/L2 > 1$, the speech rate is increased by the same factor.

In the speech encoding application where the output speech tries to mimic the input speech, both frame sizes are identical (i.e., $L2 = L1$). For other applications, however, they can be different. For example, the stored speech can be read out at a faster rate to reduce reading time. In this case, $L2 < L1$. Since the sped-up speech is not distorted, it can be easily understood even when it is played back as much as 50% faster.

Figure 11 shows that resonant frequencies and pitch are not affected by changes in speech utterance rate. A vocoder can accomplish the same feat but only with synthetic speech. We have accomplished pitch and speech rate decoupling by operating directly on the raw speech waveform.

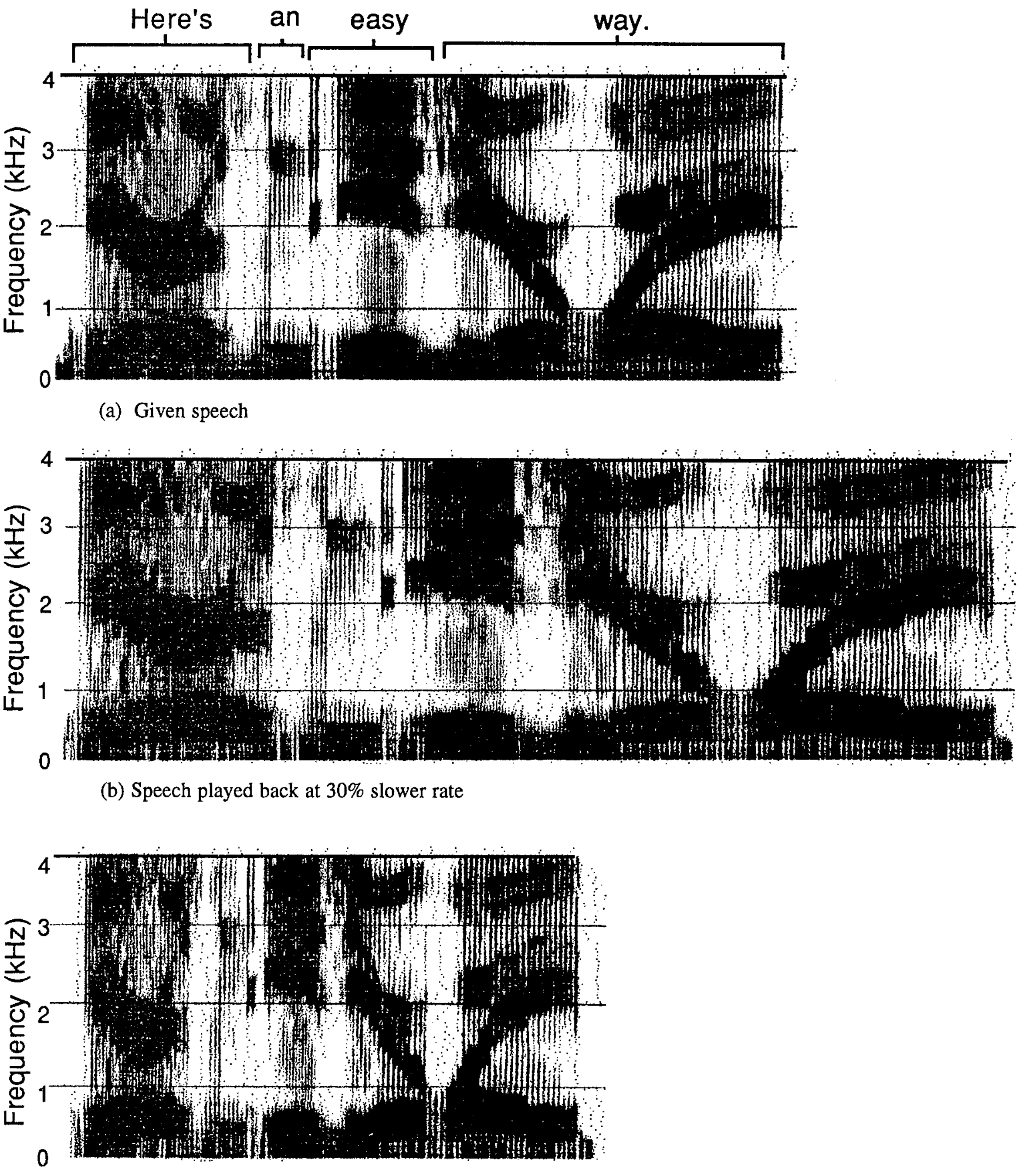
Pitch Period Modification

In speech synthesis, pitch period modification is often needed to produce appropriate intonation curves for machine-generated speech. In the vocoder model, pitch is altered by simply changing the pitch parameter. In our new speech model, the pitch synchronously-segmented waveform is expanded or compressed in accordance with the new pitch period. Altering the pitch period does not affect the speech rate, but resonant frequencies change in proportion to the pitch. As we know, high-pitched female voices have higher resonant frequencies than low-pitched male voices for the same vowel. Thus, coupling pitch frequency and resonant frequencies is beneficial.

An indispensable tool for altering pitch is the waveform expansion/compression algorithm, which is often known as the interpolation formula:

$$x(t) = \sum_{n=\infty}^{\infty} x(nt_s) \left\{ \frac{\sin\{(\pi / t_s)(t - nt_s)\}}{(\pi / t_s)(t - nt_s)} \right\}, \tag{11}$$

where t_s is the speech sampling time interval. Figure 12 is an example of waveform expansion and compression. Note that when the speech waveform is properly sampled at the Nyquist rate, no information is lost. The expanded waveform can be compressed again to become the original waveform.



(a) Given speech

(b) Speech played back at 30% slower rate

(c) Speech played back at 30% faster rate

Fig. 11 — Spectrograms of raw speech, slowed-down speech, and sped-up speech. Note that the pitch period and resonant frequencies of the original speech are not affected by modifying the speech rate. Changing the utterance rate is easily accomplished using our speech analysis/synthesis technique by adjusting the ratio of the input and output frame size.

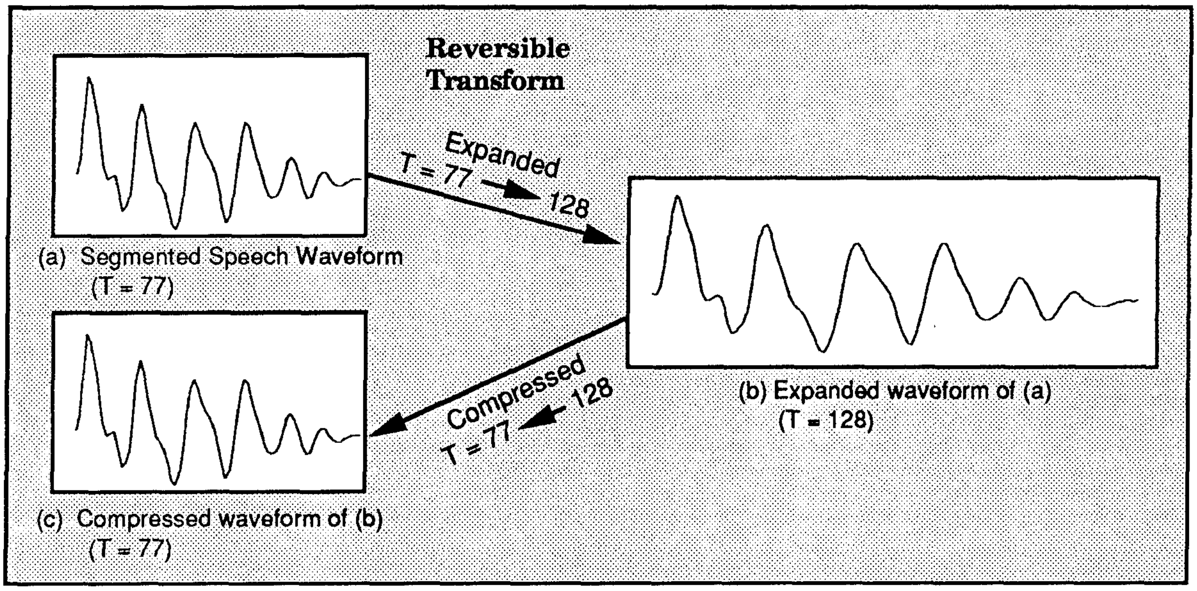

Reversible Transform
Expanded T = 77 → 128
(a) Segmented Speech Waveform (T = 77)
(b) Expanded waveform of (a) (T = 128)
Compressed T = 77 ← 128
(c) Compressed waveform of (b) (T = 77)

Fig. 12 — Examples of waveform expansion and compression. Figure 12 indicates that Eq. (11) can be used to either compress or expand the given waveform. Figure 12(a) is a segmented speech waveform where the time duration is 77 samples. Figure 12(b) is the expanded waveform of Fig. 12(a) where the time duration is increased to 128 samples. Figure 12(c) is the compressed waveform of Fig. 12(b) where time duration is reduced from 128 to 77 samples.

On the other hand, if the same waveform is first compressed then expanded, the resultant waveform will not match the original waveform because time-waveform compression introduces aliasing that cannot be eliminated by expansion.

Figure 13 illustrates the effect of pitch change where a given speech sample [Fig. 13(a)] is played back with a pitch change of –30% [Fig. 13(b)] and +30% [Fig. 13(c)]. As noted, the speech rate is not affected by the change in pitch.

Resonant Frequency Modification

Some speech processing applications need to alter the voice characteristics of the input speech. We list a few examples:

- *Disguise speaker identity*: This application is for preventing recognition of the speaker over the telephone. A simple form of frequency warping, mainly in the low-frequency region (first formant frequency region), can accomplish this. Experience shows that a small shift in the first formant frequency region (below 1 kHz) markedly alters the vocal timbre. The entire operation can be performed in real time from live speech.

- *Improve speech*: The trajectories of estimated parameters are smoother than trajectories of unprocessed speech because of the averaging process involved in parameter estimation. In other words, estimated parameters often rise at a slower rate, and their dynamic ranges are limited. Proper compensation of parameter trajectories leads to speech improvement. Researchers have noted that this is a viable research issue [10].

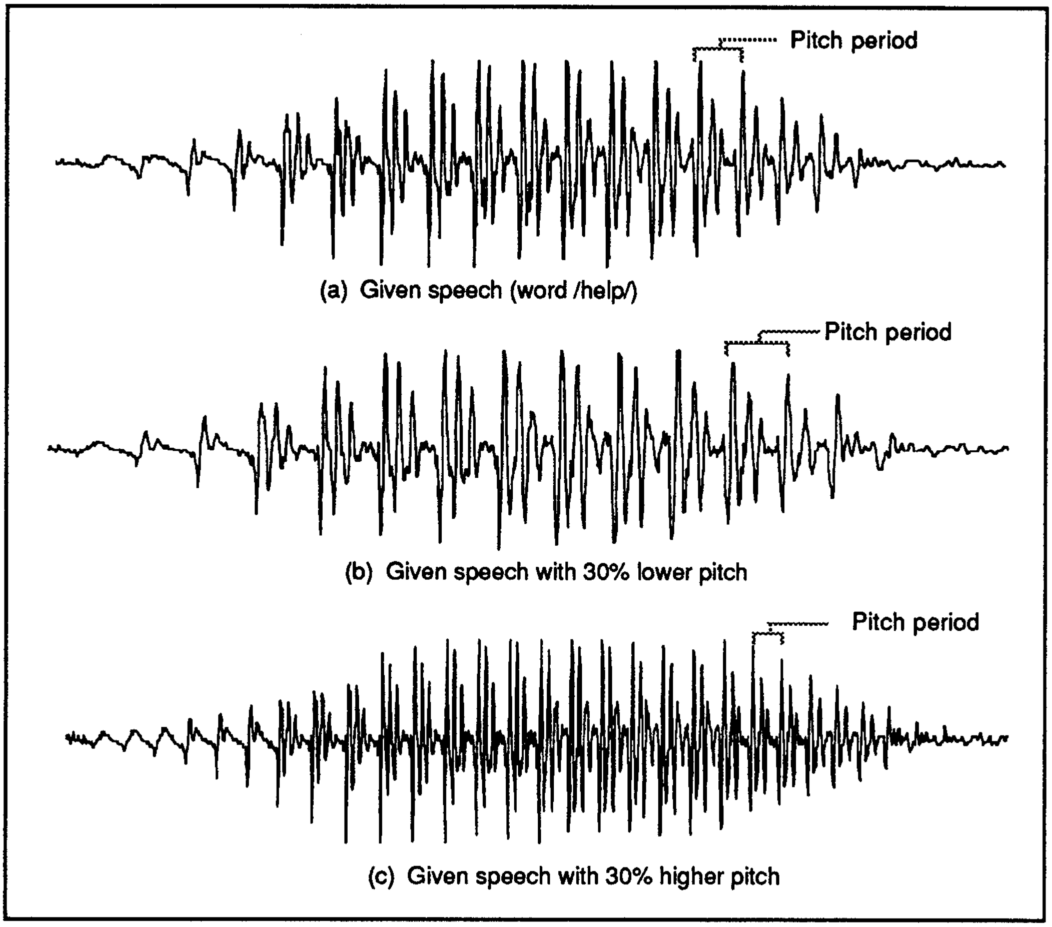


Fig. 13 — Speech waveforms with the original, lowered, and raised pitches. As noted, pitch change does not affect speech rate (i.e., all three speech waveforms have the same duration) although pitch period is either compressed or expanded.

Resonant frequency can be modified by the following four steps.

Step 1: Obtain the amplitude spectrum

In this application, transform size equals pitch period; hence, it is an even or odd number in the 20 to 120 range, not necessarily a binary-related number. Although the fast Fourier transform can be made for an integer number of time samples, the discrete Fourier transform is simpler for transforming the pitch-synchronously segmented speech waveforms. Let the sampling time interval be denoted by t_s ($< .5/B$ where B is the upper cutoff frequency of the signal in Hz), and the total number of time samples denoted by N (which may be either an even or an odd integer). Spectral components are computed for multiples of the frequency step ($f_s = 1/Nt_s$ Hz), and the total number of frequency steps is K. The total number of frequency steps in terms of the total number of time samples is

$$K = \begin{cases} (N/2)+1 & \text{if } N \text{ is even} \\ (N+1)/2 & \text{if } N \text{ is odd} \end{cases} . \qquad (12)$$

The forward transform generates the amplitude and phase spectral components. The kth amplitude spectral component $A(kf_s)$ is expressed by

$$A(kf_s) = (1 / N) [R(kf_s)^2 + X(kf_s)^2]^{1/2} \qquad k = 0, 1, 2, \ldots, K, \tag{13}$$

where

$$R(kf_s) = \sum_{n=0}^{N-1} x(nt_s) \cos(2\pi nk / N) \qquad \text{(real part)}, \tag{14}$$

and

$$X(kf_s) = \sum_{n=0}^{N-1} x(nt_s) \sin(2\pi nk / N) \qquad \text{(imaginary part)}. \tag{15}$$

The kth phase spectral component $f(kf_s)$ is expressed by

$$f(kf_s) = -\tan^{-1} \left\{ \frac{\sum_{n=0}^{N-1} x(nt_s) \sin(2\pi nk / N)}{\sum_{n=0}^{N-1} x(nt_s) \cos(2\pi nk / N)} \right\}. \tag{16}$$

This phase spectrum, however, is not used in speech generation when the amplitude spectrum is modified, to be discussed later.

Step 2: Modify the amplitude spectrum

Spectral modification is best accomplished in the amplitude spectrum because we are familiar with speech spectra and their sound effects [11]. The exact form of amplitude spectral modification varies, depending on the application, but a common spectral modification is to shift the first resonant frequency to disguise speaker identity (Fig. 14) or change the overall spectral tilt to improve the presence of speech sound.

Step 3: Compute the phase spectrum from the modified amplitude spectrum

Once the amplitude spectrum is modified, the phase spectrum expressed by Eq. (13) is no longer valid. We must derive the phase spectrum corresponding to the modified amplitude spectrum. Under the condition of causality (i.e., the signal exists only for $t \geq 0$), the phase spectrum is related to the amplitude spectrum through the Hilbert transform [12]. We discuss the steps required to derive phase spectrum from modified amplitude spectrum in a later section related to speech encoding.

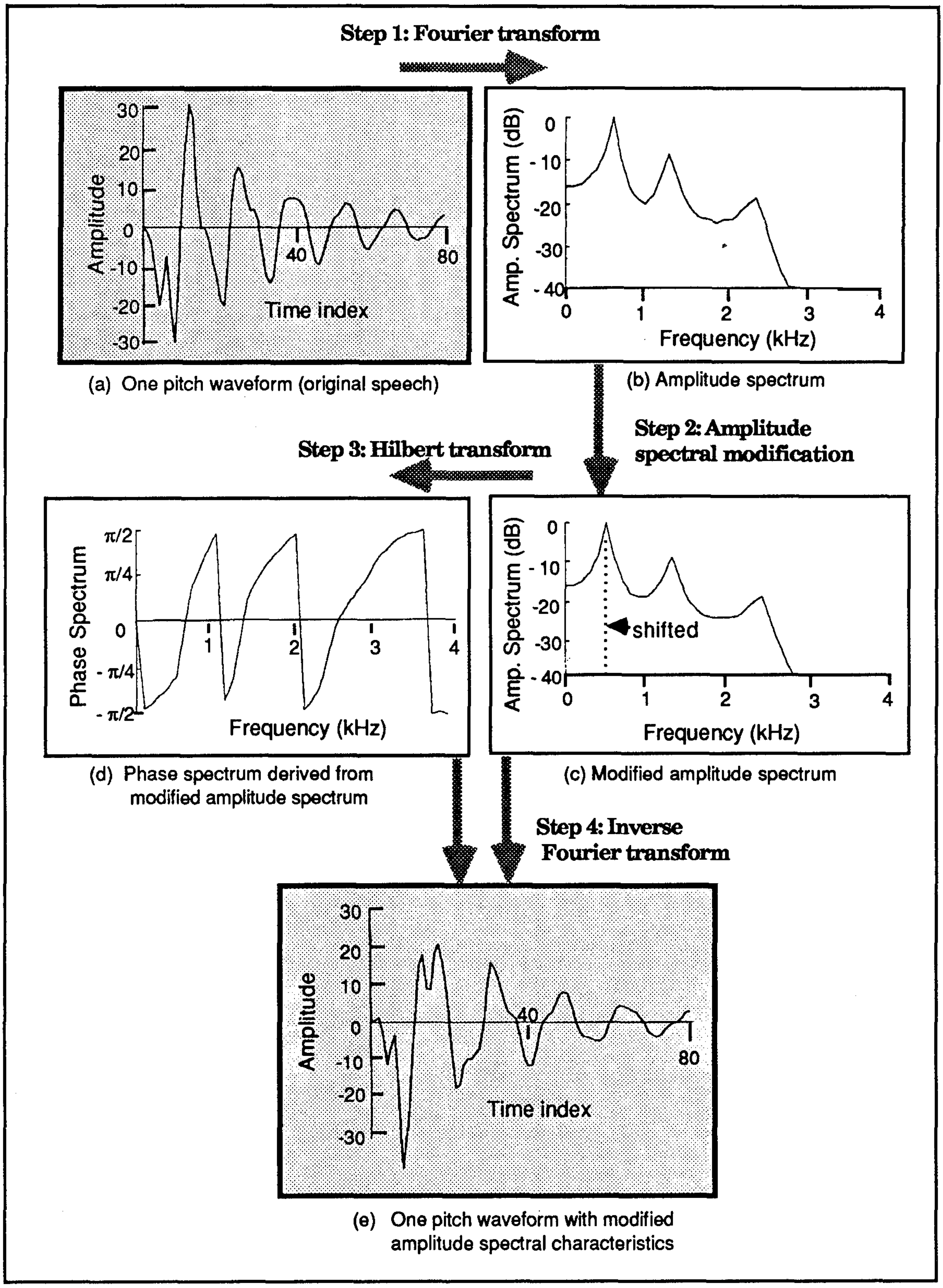


Fig. 14 — Waveforms associated with steps required to alter the amplitude spectral characteristics of the segmented speech waveform. In this example, the first resonant frequency is shifted down by 150 Hz (Fig. 14(c)). Note that the output speech waveform (Fig. 14(e)) is significantly different from the input speech waveform (Fig. 14(a)).

Step 4: Inverse Fourier transform to generate modified speech waveform

To perform the inverse Fourier transform, the end-point amplitude spectral components get a factor of 0.5 because the summations in Eqs. (14) and (15) involve only positive frequencies. Thus,

$$A(0) \rightarrow 0.5A(0) \qquad \text{if } N \text{ is even or odd} \tag{17}$$

$$A(K) \rightarrow 0.5A(K) \qquad \text{if } N \text{ is even.} \tag{18}$$

This end-point correction is mandatory; otherwise, the back-to-back transform will not be transparent (Table 2). The output time sample, obtained from the amplitude and phase spectral components, is

$$x(nt_s) = \sum_{k=0}^{K-1} A(kf_s)\cos[(2\pi nk/N) + f(kf_s)] \tag{19}$$

where $n = 0, 1, 2, ..., N$-1.

Table 2 — Back-to-Back Fourier and Inverse Fourier Transforms

(a) N = 25 (odd)

Index	DFT Input	DFT Output		IDFT Output
	Time Sequence	Amplitude Spectrum	Phase Spectrum	Time Sequence
0	1.00	6.50	0.00	1.00
1	2.00	3.99	1.70	2.00
2	3.00	2.01	1.82	3.00
3	4.00	1.36	1.95	4.00
4	5.00	1.04	2.07	5.00
5	6.00	0.85	2.20	6.00
6	7.00	0.73	2.32	7.00
7	8.00	0.65	2.45	8.00
8	9.00	0.59	2.58	9.00
9	10.00	0.55	2.70	10.00
10	11.00	0.53	2.83	11.00
11	12.00	0.51	2.95	12.00
12	13.00	0.50	3.08	13.00
13	14.00			14.00
14	15.00			15.00
15	16.00			16.00
16	17.00			17.00
17	18.00			18.00
18	19.00			19.00
19	20.00			20.00
20	21.00			21.00
21	22.00			22.00
22	23.00			23.00
23	24.00			24.00
24	25.00			25.00

Input equals output

(b) N = 24 (even)

Index	DFT Input	DFT Output		IDFT Output
	Time Sequence	Amplitude Spectrum	Phase Spectrum	Time Sequence
0	1.00	6.25	0.00	1.00
1	2.00	3.83	1.70	2.00
2	3.00	1.93	1.83	3.00
3	4.00	1.31	1.96	4.00
4	5.00	1.00	2.09	5.00
5	6.00	0.82	2.23	6.00
6	7.00	0.71	2.36	7.00
7	8.00	0.63	2.49	8.00
8	9.00	0.58	2.62	9.00
9	10.00	0.54	2.75	10.00
10	11.00	0.52	2.88	11.00
11	12.00	0.50	3.01	12.00
12	13.00	0.25	3.14	13.00
13	14.00			14.00
14	15.00			15.00
15	16.00			16.00
16	17.00			17.00
17	18.00			18.00
18	19.00			19.00
19	20.00			20.00
20	21.00			21.00
21	22.00			22.00
22	23.00			23.00
23	24.00			24.00

Input equals output

Table 2 illustrates back-to-back Fourier transforms for cases where the number of input time samples is odd (N = 25) and even (N = 24). The phase spectral components circled in Table 2 contain no information (i.e., always 0 or π radians). Thus, the number of spectral values equals the number of time samples. In other words, the number of parameters is the same before and after the transform.

Note that data identified by circles contain no useful spectral information because they are independent of the input. Thus, the number of spectral components is N (whether N is even or odd), which is identical to the number of input time samples. As noted, the back-to-back Fourier transform is transparent (i.e., the input time series equals the output time series) in the absence of spectral quantization.

Figure 14 illustrates the waveform involved in each step for modifying the amplitude spectral characteristics of speech. Note the significant changes in the output speech waveform caused by the shift of the first resonant frequency by only 150 Hz.

APPLICATION TO SPEECH ENCODING

Another significant application of the new analysis/synthesis technique is for encoding speech at low data rates, 2400 b/s or less. A data rate of 2400 b/s is critically important for the Navy because Naval communication depends on narrowband links such as high-frequency (HF), and MILSTAR and FLTSATCOM satellites. Even if wideband links become available in the future, a data rate of 2400 b/s cannot be discarded because the low data rate is amenable to incorporating an anti-jam feature into voice communication (as in MILSTAR, where the data rate is 2400 b/s). Also, a lower data rate requires a lower transmission power to reach the same distance.

In the 1950s and 60s, the 2400-b/s voice encoders were channel vocoders based on spectral characterization of speech over 16 to 19 discrete frequency bands. In the early 1980s, channel vocoders were replaced by LPCs. The U.S. conducted numerous tests which indicated that LPC was better than other voice processors in terms of speech intelligibility, speech quality, acoustic noise immunity, bit error-performance, and tandem performance. Tactical communicators welcomed the 2400-b/s LPC when it first appeared, but current users have forgotten the improvements that LPC offered over the old channel vocoders, and they tend to compare LPC unfavorably with the familiar telephone.

Before describing our new 2400-b/s voice encoder, we examine the known limitations of LPC, and we discuss the reasons behind these limitations. At the same time, we point out why our new voice technique is free from the causes responsible for the performance degradation in the current LPC.

Limitations of Current 2400-b/s LPC

There are two notable weaknesses in the current 2400-b/s LPC:

- familiar voices cannot be recognized easily, and
- speech intelligibility of female voices is significantly lower than that of male voices.

We examine these weaknesses in the following.

Poor Speaker Recognizability

One of the limitations of the 2400-b/s LPC is a lack of speaker recognizability. Astrid Schmidt-Nielsen of NRL tested speaker recognition with 24 coworkers [5]. They listened to both 2400-b/s LPC processed speech and unprocessed speech and tried to identify each talker from a group of about 40 people working in the same office. Unprocessed voices were correctly identified 88% of the time, whereas the same voices through the LPC system were correctly identified only 69% of the time. Schmidt-Nielsen is currently planning to test speaker recognition by using current 2400-b/s LPC.

Stephanie Everett of NRL tested the ability of automatic speaker recognition algorithms to recognize speakers by using encoded speech as input. Currently, automatic speaker recognizers are used for controlling access to secure telephones or restricted areas. In this evaluation, four types of vocoded speech encoders were used [13]:

- 2400-b/s LPC
- 9600-b/s Residual-Excited LPC (RELP) [14]
- 16000-b/s RELP [14]
- 64000-b/s Pulse Code Modulator (PCM).

From this research, we can relate speech intelligibility and automatic speaker recognition scores. It is significant to note that the automatic speaker recognition scores are nearly directly proportional to the vocoded speech intelligibility (Fig. 15). Thus, improving the intelligibility of encoded speech should improve speaker recognizability.

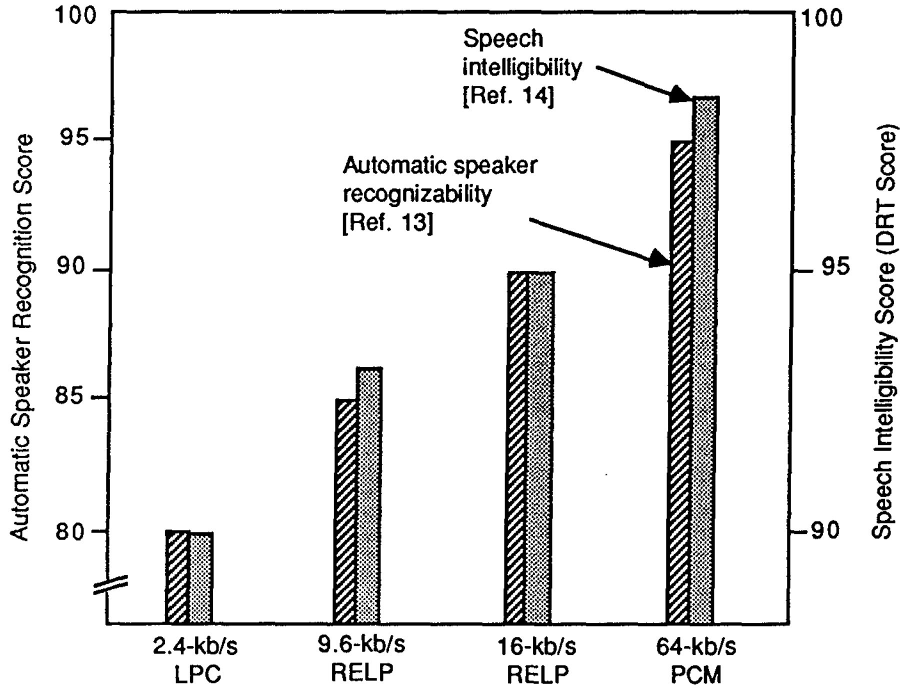


Fig. 15 — Comparison between automatic speaker recognizability and speech intelligibility for four different speech encoders. Surprisingly, they are highly correlated. As expected, the automatic speaker recognizability of the 2.4-b/s LPC is the worst. To improve speaker recognizability, we must improve speech intelligibility.

Although the performance of speaker recognition by human subjects may be different from that of computers, we would be rather surprised if the conclusion is significantly different from that shown in Fig. 15. Astrid Schmidt-Nielsen is currently investigating this issue at NRL by using more recent voice encoders.

Poor Speech Intelligibility for Female Speakers

Another limitation of the 2400-b/s LPC is that the intelligibility of female voices is not as good as that of male voices. We recently examined DRT scores collected by the DoD Voice Processor Consortium under the direction of Thomas Tremain [15]. We note that the average intelligibility of female speakers is 5.1 points below that of male speakers (Table 3). Ten years ago, we examined the same issues using older 2400-b/s LPCs. The result was essentially the same. A DRT loss of 5 points is equivalent to speech degradation caused by random bit errors as high as 2 percent.

Table 3 — Speech Intelligibility of Male and Female Voices over 2400-b/s LPC

Test Condition				DRT Dcore		
Microphone	Noise	Bit Error	Tandem	Male	Female	Diff.
Dynamic	-	-	-	92.9	86.2	- 6.7
Carbon	-	-	-	90.2	85.7	- 4.5
Dynamic	-	1% Random	-	88.9	85.0	- 3.9
	-	2% Random	-	88.3	82.9	- 5.4
	-	5% Random	-	84.1	75.5	- 8.6
	-	1% Block	-	92.6	86.5	- 6.1
	-	5% Block	-	89.5	85.0	- 4.5
Noise-Cancelling	Office	-	-	88.8	84.9	- 3.9
	Ship	-	-	86.9	78.5	- 8.4
	E4B	-	-	86.9	84.6	- 2.3
	P3C	-	-	83.5	81.7	- 1.8
	Tank	-	-	83.8	75.9	- 7.9
Dynamic	-	-	Self	89.5	83.9	- 5.6
	-	-	Into APC16	87.5	82.6	- 4.9
	-	-	From APC16	84.2	79.7	- 4.5
	-	-	Into CVSD16	85.3	83.7	- 1.6
	-	-	From CVSD16	84.9	80.2	- 4.7
	-	-	Into CVSD32	89.9	82.9	- 7.0
	-	-	From CVSD32	87.6	82.4	- 5.2

Average - 5.1

Reasons for Poor Performance

The reason for the poor speech intelligibility of the female voice lies with the LPC principle itself. Linear predictive analysis is based on the principle that a speech sample is represented by a linear combination of past samples:

$$x_i = \sum_{n=1}^{N} \alpha_n x_{i-n} + \varepsilon_i \qquad i = N+1, N+2, \ldots, I, \tag{20}$$

where x_i is the ith speech sample, α_n is the nth prediction coefficient, and ε_i is the ith prediction residual. Equation (20) is a good representation of speech within each cycle where the speech waveform is continuous (i.e., glottis excitation is absent). At each pitch epoch, the glottis excitation is renewed and the speech waveform becomes discontinuous. Hence, a speech sample after a pitch epoch cannot be expressed in terms of speech samples prior to the pitch epoch (Fig. 16). However, we have been doing this since the linear predictive analysis became popular in the early 1970s.

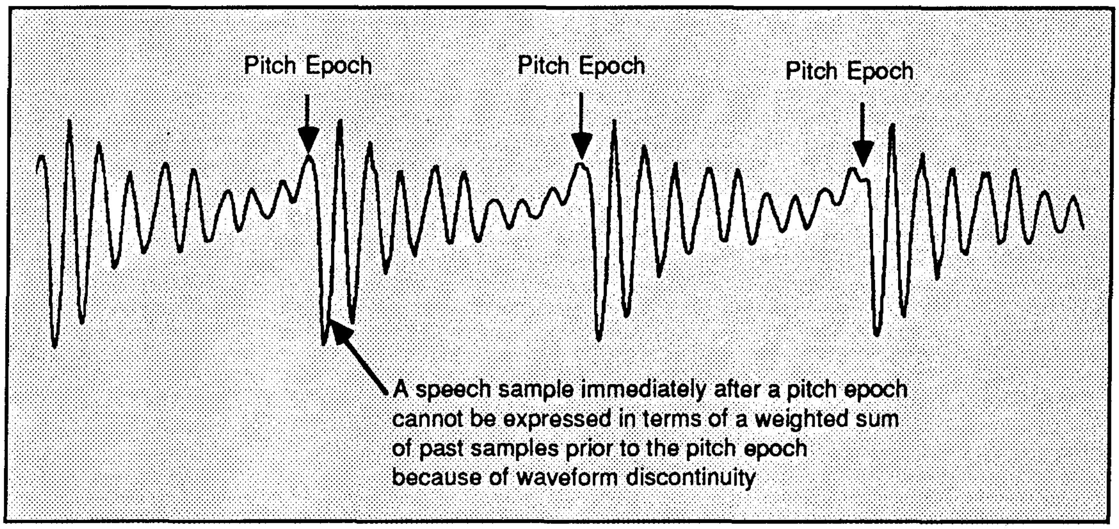


Fig. 16 — Voiced speech waveform. The principle of linear prediction expressed by Eq. (20) is invalid in the vicinity of pitch epochs because the speech waveform is discontinuous. The presence of pitch modulation within the LPC analysis window (which we call the *intra-frame pitch interference*) degrades the quality of LPC analysis results. As illustrated in Fig. 17, the LPC spectrum is a poor approximation of the speech spectral envelope when the pitch is high.

This least-squares principle is valid as long as all of the speech samples appearing in Eq. (20) come from the same pitch cycle. This situation rarely occurs because the LPC analysis window size is typically 130 to 160 samples, whereas the average pitch period is as small as 20 samples for a female voice and 80 for a male voice. In each LPC analysis window, there are more pitch epochs for a female voice than for a male voice. As a result, the female voice is not characterized as well as the male voice. In other words, the LPC spectrum is more distorted for the female voice than for the male voice. To illustrate the effect of the pitch interference on the LPC spectrum, we computed the LPC spectrum in two different ways:

(1) Using the full set of prediction equations defined by Eq. (20): The LPC coefficients are computed by using all of the speech samples within the analysis window. Parameters used for

Eq. (20) are $N = 10$ coefficients and analysis window size is $I = 130$. The 10 prediction coefficients are obtained from the solution of 120 simultaneous linear equations stemming from Eq. (20). This is the conventional approach for computing LPC coefficients, but they contain the effect of pitch interference. Figure 17(a) shows the resultant LPC spectrum. As noted, the second and third resonant frequencies are not clearly distinguishable.

(2) Using selected prediction equations with smaller prediction errors: The LPC coefficients are computed by using only the equations that produce acceptably small prediction errors (i.e., less than twice the standard of deviation). Because we have 120 equations with 10 unknowns, we could eliminate some of the equations (approximately 10 equations per pitch epoch) that contribute to a large prediction error. Figure 17(b) shows the resultant LPC spectrum. As noted, the second and third resonant frequencies are more clearly identifiable.

If the pitch is high, LPC spectral deterioration is not limited to just obscuring resonant frequencies. The consequence is even more disastrous. Figure 18(b) shows that the estimated LPC spectrum of high-pitched voices tends to follow the pitch harmonics, rather than the speech spectral envelope. The resultant synthetic speech becomes reverberant and less intelligible because pitch harmonics are supplied not only by the excitation signal (as they should be) but also by the vocal tract filter as the result of pitch interference.

New 2400-b/s Speech Encoder

Existing vocoders (including the 2400-b/s LPC) characterize the speech waveform in terms of parameters associated with the electric analog of the vocal tract shown in Fig. 1. In .contrast, we characterize the speech waveform itself. There are two significant advantages to our approach:

- Elimination of intraframe pitch interference: The intra-frame pitch interference is a result of each analysis window containing the speech waveform of multiple pitch cycles. The consequence of having pitch interference was discussed in the preceding section. In the new speech encoding approach, the intraframe pitch interference is completely eliminated by transmitting the speech waveform defined in a *single* pitch cycle.

- Elimination of interframe pitch interference: The voiced speech waveform (i.e., vowels) is periodic at the rate of the fundamental pitch frequency. Thus, short-term averaged speech parameters (such as the rms value, autocorrelation function, LPC coefficients, etc.) tend to fluctuate at the pitch rate. From the early days of vocoding technology, pitch-synchronous speech analysis was emphasized to minimize flutters in the synthesized speech. By placing the analysis window in synchronization with the pitch cycles, the interframe pitch interference (i.e., modulation caused by pitch-to-pitch variations) is minimized. In the new speech encoding approach, interframe pitch interference is completely eliminated because the analysis window is always placed at the pitch epoch.

Elimination of these two types of pitch interference enhances the speech intelligibility at comparable data rates.

Overview

Frame size is the time interval between parameter or waveform updates. Our frame size is 180 samples (or 22.5 ms), which is identical to that of the current 2400-b/s LPC. However, we transmit certain parameters more than once per frame. This aspect is discussed in the next section. For a data rate of 2400 b/s, 54 bits are allocated to each frame (Table 4).

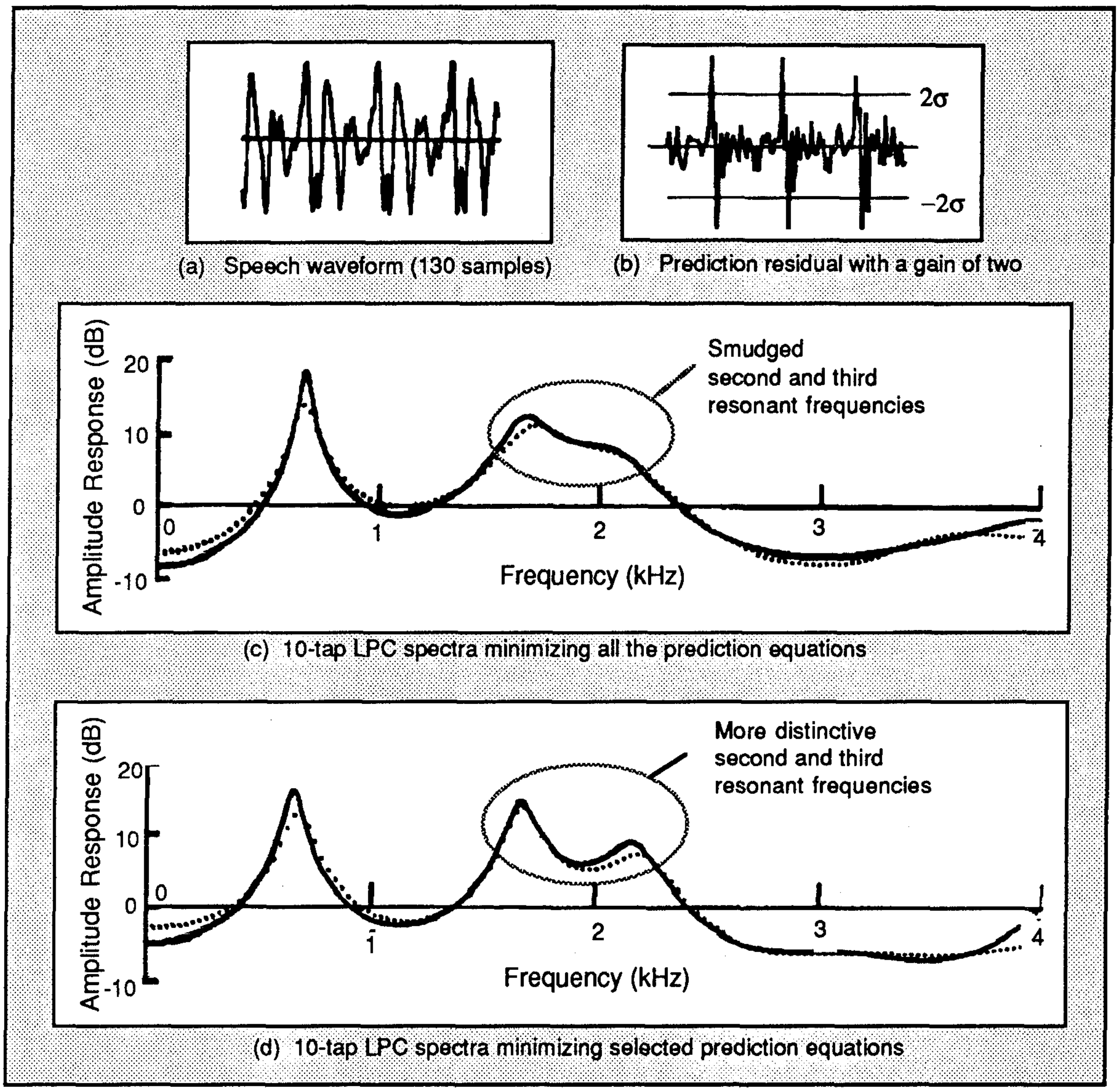
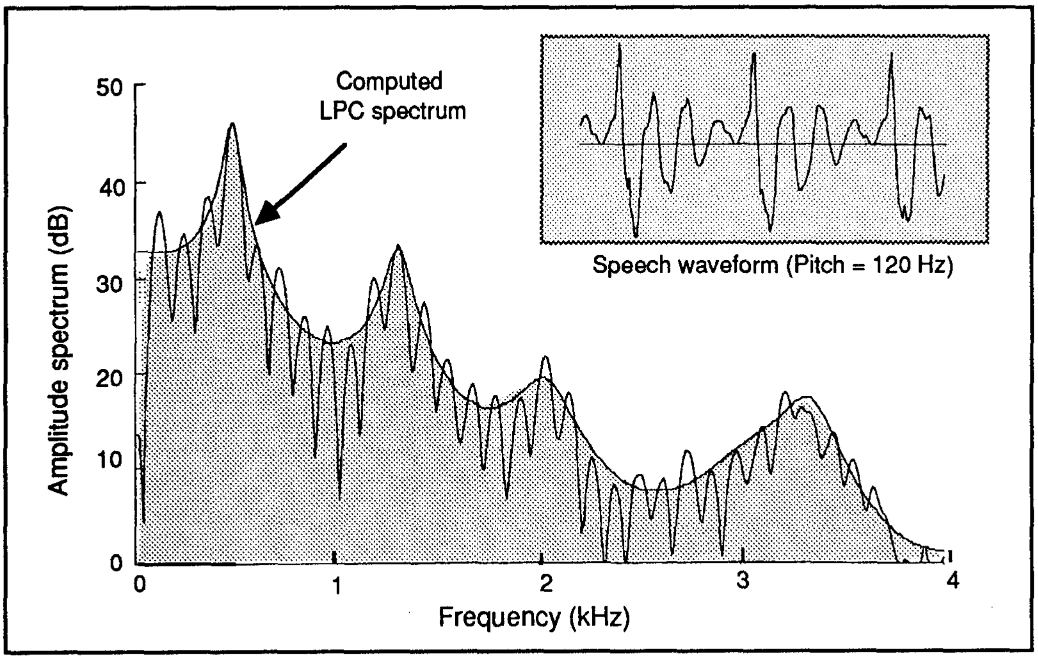
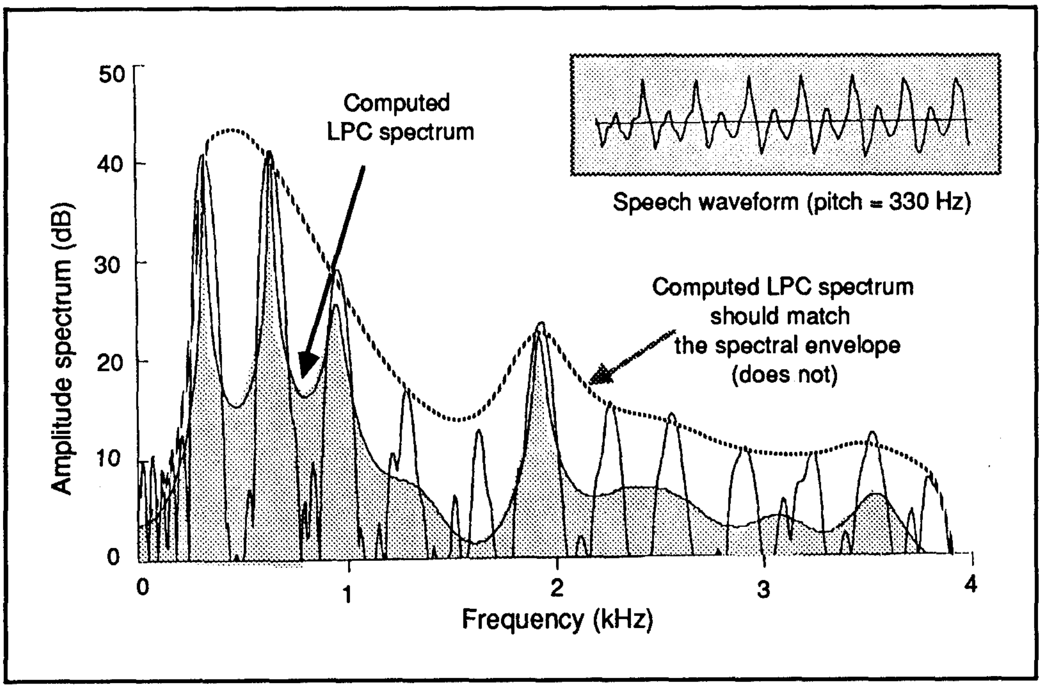


Fig. 17 — Speech waveform and LPC spectra estimated by linear prediction method. Figures 17(a) and (b) are the speech waveform and the prediction residual, respectively. As noted, prediction residual (prediction error) is larger near the pitch epoch where the prediction principle breaks down. Figure 17(c) is the 10-tap LPC spectrum using all the prediction equations (i.e., 130-10 = 120 equations). Figure 17(d) is also a 10-tap LPC spectrum from the same speech waveform, but the prediction equations contributing to large residuals (greater than ±2σ) are excluded. As noted, the second and third resonant frequencies are better defined. Dotted lines in Figs. 17(c) and (d) are LPC spectra with coefficients quantized for 2400 b/s. The effect of parameter quantization is negligible in comparison with the effect of pitch interference.



(a) Vowel from low-pitch male voice



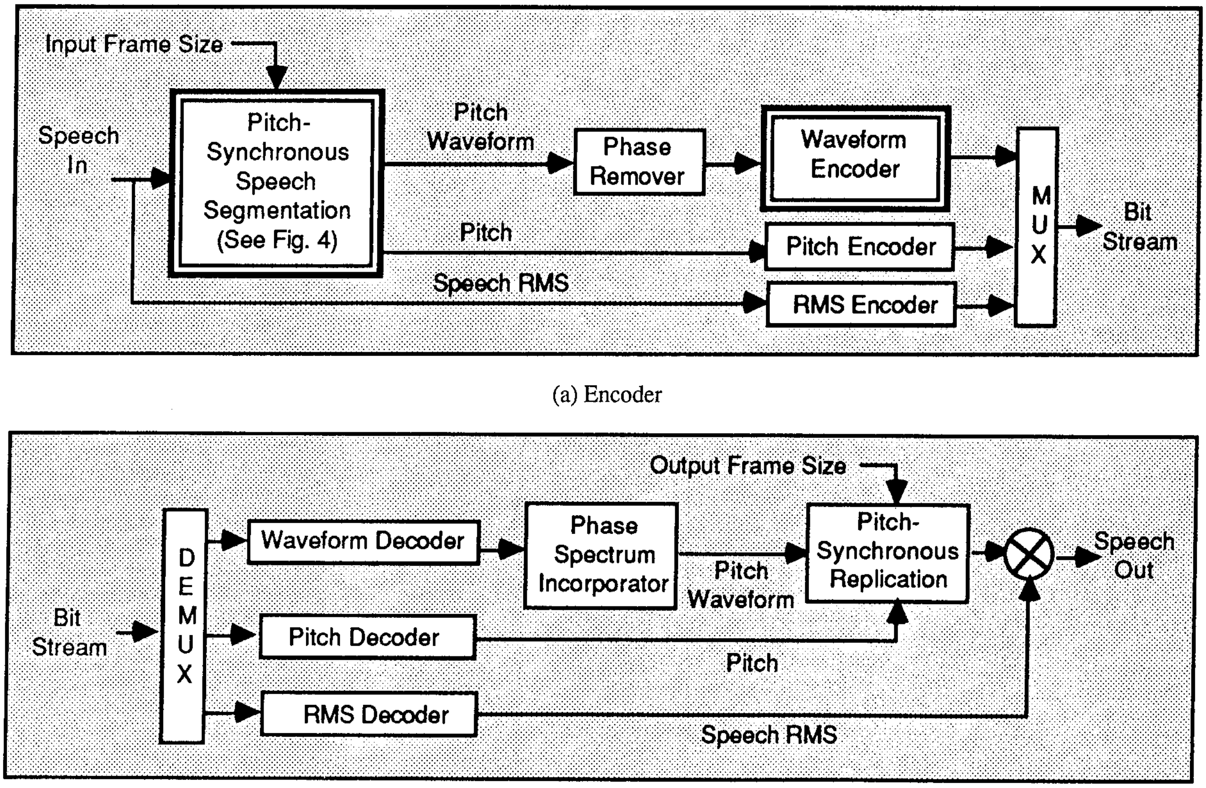
(b) Vowel from high-pitch female voice

Fig. 18 — Two examples of LPC speech spectra and their estimated spectral envelopes. Figure 18(a) shows that when pitch is low, the LPC spectrum follows the speech spectral envelope well. As a result, LPC-synthesized speech of a male voice sounds reasonably good. In contrast, Fig. 18(b) shows that when pitch is high (i.e., pitch harmonics are sparsely spaced), the LPC spectrum tends to follow the pitch harmonics rather than the spectral envelope. The synthesized speech will sound reverberant because pitch harmonics are supplied by both the excitation signal and the LPC coefficients. As a result, intelligibility of the female voice is degraded.

Table 4 — General Information About the New 2400-b/s Speech Encoder

Speech sampling rate	8 kHz ± 0.1%
Data rate	2400 b/s
Frame size	180 samples or 22.5 ms
Frame rate	44.444 Hz
No. of bits per frame	54 bits

The new speech encoder is a hybrid between waveform vocoders such as pulse code modulation (PCM) and pitch-excited vocoders that renew the excitation at the average pitch rate (2400-b/s LPC). The new speech encoder is similar to PCM because it transmits one cycle of the speech waveform during each frame. On the other hand, the new speech encoder is similar to the pitch-excited vocoder because the pitch waveform is repeated at the pitch rate. Figure 19 is a block diagram of the new speech encoder. It shows that three different speech parameters are encoded: pitch period, speech rms, and the pitch-synchronously segmented speech waveform. Table 5 lists bit allocations for these three parameters. The reasons for this bit allocation are justified in subsequent sections.



(a) Encoder

(b) Decoder

Fig. 19 — The new 2400-b/s speech encoder. Two critical elements are indicated by double-lined boxes, one of which is "Pitch-Synchronous Speech Segmentation" detailed in Fig. 4.

Table 5 — Number of Bits per Frame Allocated for Each Parameter

	Pitch Period < 40		40 ≤ Pitch Period ≤ 80		Pitch Period > 80	
Synchronization	Once per frame	1 bits	Once per frame	1 bits	Once per frame	1 bits
Pitch	Once per frame	7	Once per frame	7	Once per frame	7
Speech rms	Three times per frame	12*	Twice per frame	9*	Once per frame	5
Pitch waveform	Thee times per frame	34*	Twice per frame	36*	Once per frame	41
* Vector quantization	TOTAL:	54 bits		54 bits		54 bits

Repeating the waveform more than two to three times per frame adversely affects the speech intelligibility. Therefore, pitch waveforms and speech rms values are encoded two to three time per frame, depending on the pitch value (Table 5). We discuss a method of quantizing each parameter.

Synchronization Bit

To acquire and maintain frame synchronization, we use an alternating "1" and "0" in each frame.

Pitch Period Quantization

The pitch range is from 16 to 133 for all consecutive pitch periods (7-bit quantity). Thus, the allowable fundamental pitch frequency is from 500 Hz to 60.15 Hz, which is not significantly different from that of the 2400-b/s LPC. A pitch value and its corresponding code are related by

$$\tau_{\text{code}}(i) = \tau(i) - 16 \tag{21}$$

and

$$\tau(i) = \tau_{\text{code}}(i) + 16, \tag{22}$$

where $\tau(i)$ is the *i*th pitch period ($16 \leq \tau(i) \leq 133$), and $\tau_{\text{code}}(i)$ is the corresponding pitch code ($0 \leq \tau_{\text{code}}(i) \leq 127$). The pitch period is transmitted once per frame.

Speech rms Quantization

The speech rms controls the loudness of the synthesized speech. As indicated in Table 5, speech rms value is transmitted once, twice, or three times, depending on pitch period.

Case 1: Pitch Period > 80: If the pitch period is greater than 80 (mainly low-pitch male voices), only one rms value is transmitted per frame. In this case, the speech rms value is semilogarithmically quantized into 5 bits by the same quantization table used in the 2400-b/s LPC [7].

Case 2: 40 ≤ Pitch Period ≤ 80: If the pitch period is greater than or equal to 40 but less than or equal to 80, two half-frame rms values are transmitted in each frame. We note that two consecutive speech rms values separated by a short time interval (less than 100 ms) are correlated (Fig. 20) because we cannot alter our speaking volume suddenly. Thus, it is advantageous to encode two rms values jointly (i.e., vectorially) quantized. Initially, each rms value is logarithmically quantized into one of 26 values. Then, two logarithmically quantized rms values (denoted by A1 and A2, respectively) are jointly encoded by a two-dimensional look-up table (Table 6). According to extensive analyses of various speech samples, only 512 are significant among 676 (= 26 × 26) possible rms transitions over a

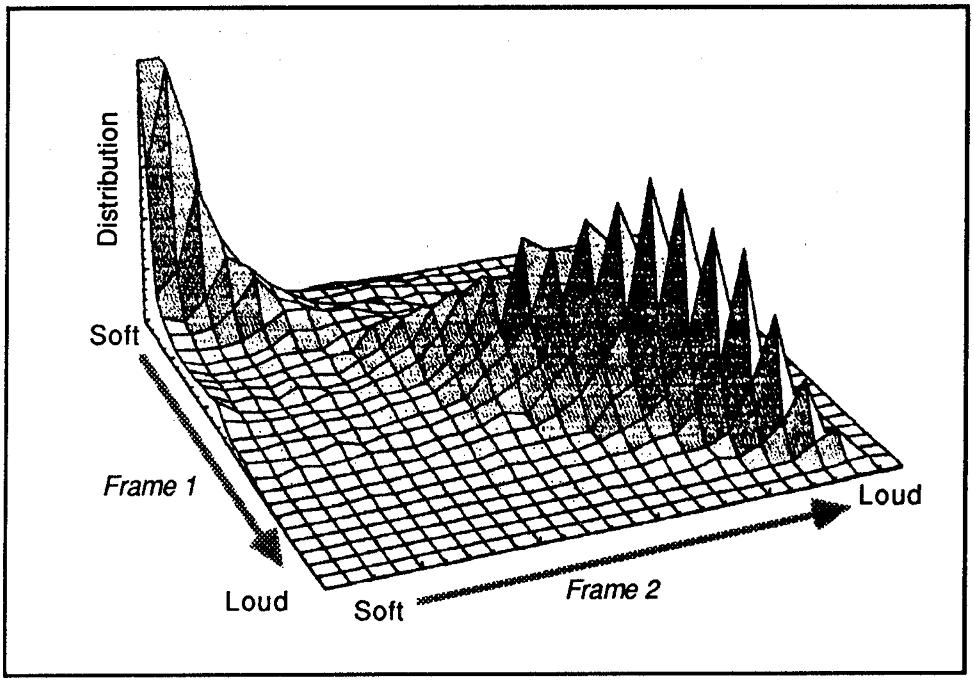
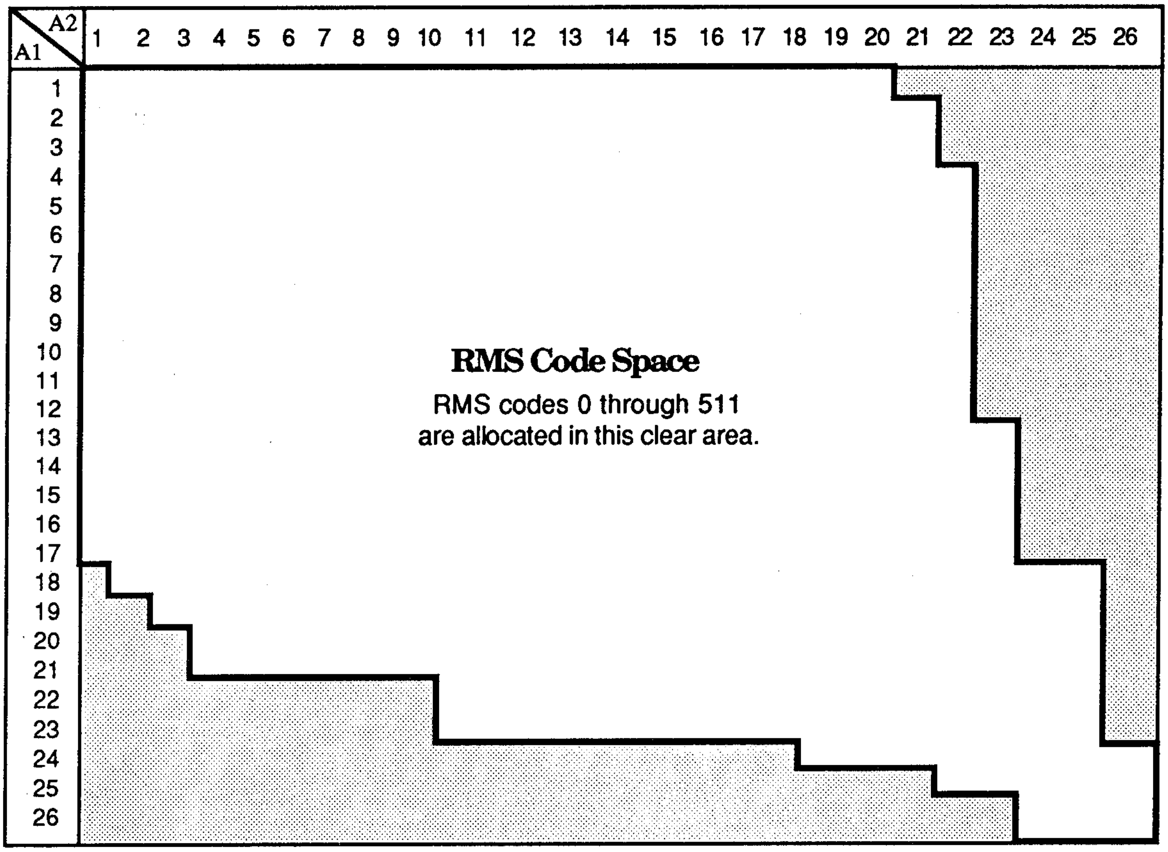


Fig. 20 — Joint distributions of two speech rms values obtained a short time apart. As noted, these two rms values are highly correlated. This figure implies that humans cannot make soft and loud sounds in quick succession.

Table 6 — Two-dimensional rms Coding/Decoding Table when 40 ≤ Pitch Period ≤ 80


A2
A1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
RMS Code Space
RMS codes 0 through 511
are allocated in this clear area.

short period time (i.e., 90 sampling time intervals). Table 6 shows that each of the allowable rms transitions is assigned a code. Thus, 9 bits are needed to encode speech rms values rather than 10 bits (two 5 bits) if the two rms values are encoded independently.

Case 3: Pitch Period < 40: When the pitch period is less than 40 (mainly female voices), we transmit the speech rms value three times per frame (Table 5). In this case, we use a 3-dimensional rms encoding table similar to that used for 2-dimensional rms encoding.

Pitch Waveform Quantization

Encoding the pitch-synchronously segmented waveform is the most critical element in the new speech encoder because it critically affects the synthesized speech quality. The pitch waveform can be encoded in the time domain or in the frequency domain. In either case, we perform the following three operations to make encoding more efficient:

1. Time normalization of the pitch waveform: To make quantization more effective, the pitch-synchronously segmented waveform is stretched to 40 if pitch period < 40, 80 if 40 ≤ pitch period ≤ 80, or 120 if pitch period > 80. If the pitch period is greater than 120, it is clamped to 120.

2. Removal of phase spectral information: The phase information is not too significant to the pitch-synchronously segmented speech waveform because the time origin is always set at the pitch epoch. However, we cannot simply discard the phase information because our ears can hear relative phase changes from one pitch waveform to next. The phase information that is adequate for regenerating speech may be obtained from the amplitude spectrum via the Hilbert transform, or a limited number of artificial phase spectra can be used (i.e., one phase spectrum for vowels and several phase spectra for fricatives and stop consonants). The use of artificial phase spectra is simpler, and is used in our system. By removing phase information, the number of pitch waveform samples is reduced to half. This is true whether the information is encoded in the frequency domain or in the time domain.

3. Pattern matching: The human voice is incapable of making certain sounds. Thus, the distribution of pitch-waveform spectra has a null space (Fig. 21). The pattern-matching process effectively excludes encoding of the speech spectra corresponding to the null space because the reference templates are generated from actual human speech.

To implement the pattern-matching process, we collected three sets of templates, one each from the three pitch ranges (pitch period < 40, 40 ≤ pitch period ≤ 80, and pitch period > 80) from the Texas Instruments - Massachusetts Institute of Technology (TIMIT) Acoustic-Phonetic Speech Data Base [16]. For each pitch range, the templates were collected through the following three steps:

Step 1: The amplitude spectrum of the first incoming pitch waveform becomes the first template, and it is stored in memory.

Step 2: The amplitude spectrum of the next incoming pitch waveform is compared with each of the stored templates. If the difference is larger than an acceptable level (i.e., an average spectral error of 2 dB), the spectrum of this pitch waveform is stored as a new template. Otherwise, it is discarded.

Step 3: Step 2 is repeated until the maximum allowable number of templates (Table 6) is reached. Actually, we collect more than the maximum number, then eliminate the least-frequently-used templates later on to meet the required maximum template size.

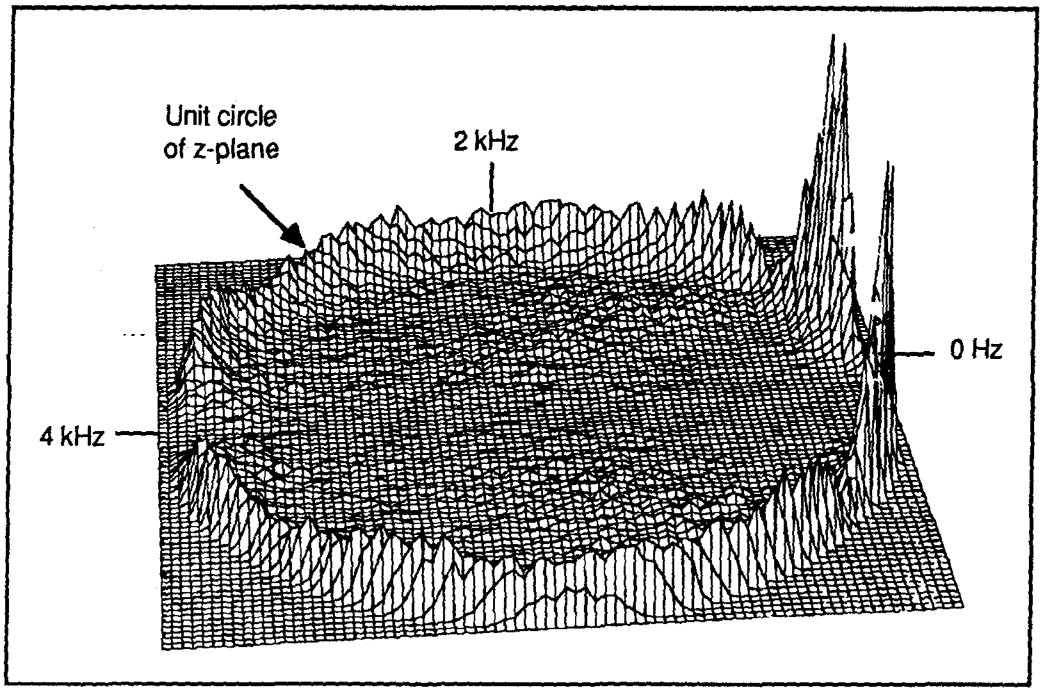


Fig. 21 — Distributions of amplitude spectra (expressed in the z-domain) of pitch waveforms. As noted, there is null space because we do not make certain sounds in our normal conversation. By excluding this null space from encoding, we achieve an efficient speech transmission.

Phase Spectrum Restoration

Amplitude and phase spectra are not generally related. If the input time sequence is causal, however, the phase and amplitude spectra are related by the Hilbert transform [12]:

$$\arg[X(e^{j\omega})] = (-1/2\pi)\, P \int_{-\pi}^{\pi} \log|X(e^{j\theta})|[\cot(\omega-\theta)/2]\, d\theta, \tag{23}$$

where $\log|X(e^{j\theta})|$ is the log amplitude spectrum of the real input sequence $x[n]$, $\arg[X(e^{j\omega})]$ is the phase spectrum of $x[n]$, and P denotes the Cauchy principle value of the integral that follows. Figure 22 compares the group delays (first derivative of the phase spectrum) computed from the amplitude spectrum by use of Eq. (23) with the original group delay computed directly from the input sequence. They agree with each other very well.

Intelligibility Scores

Any new 2400-b/s speech encoder must outperform the existing 2400-b/s LPC, especially in difficult operating conditions such as female speech and noisy speech. We processed DRTs for these two cases. The voice processors we compared were:

- 2400-b/s LPC,
- our new 2400-b/s voice encoder, and
- 4800-b/s CELP.

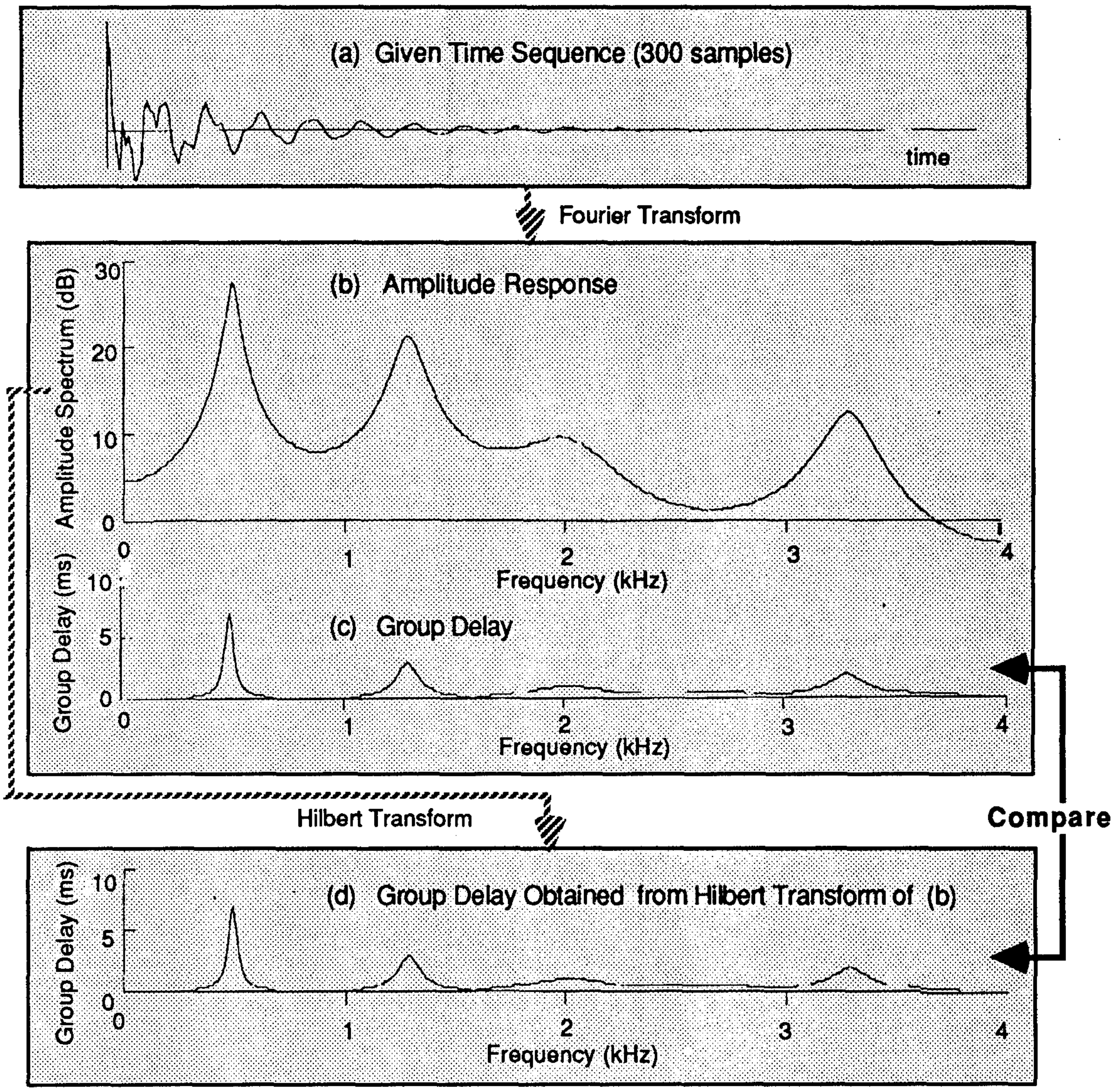
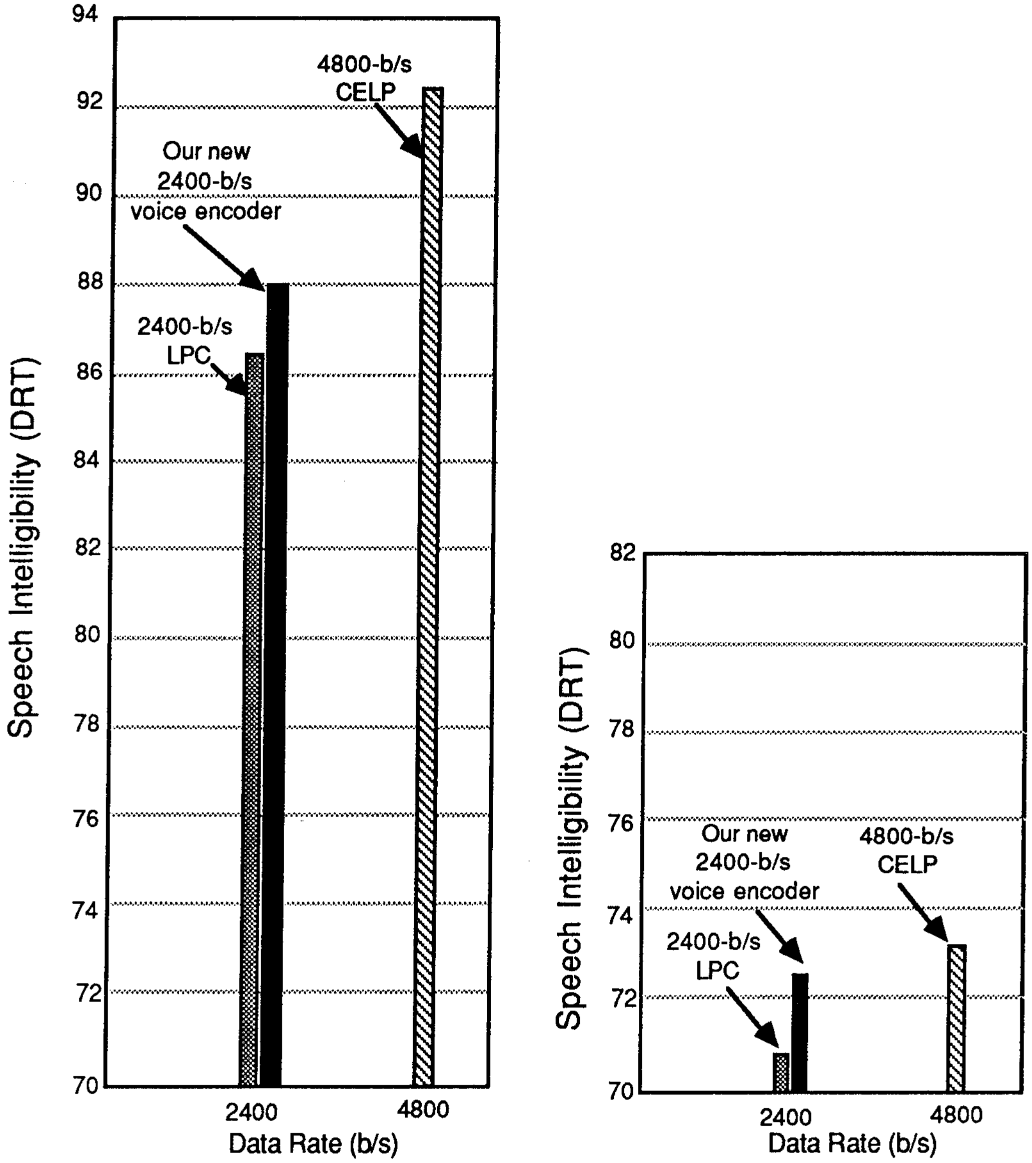


Fig. 22 — Comparison of two group delays. Time sequence (Fig. 22(a)) generates both the amplitude spectrum (Fig. 22(b)) and group delay (Fig. 22(c)) via the Fourier transform. Since the time sequence is causal (i.e., $x[n] = 0$ for $n < 0$), the Hilbert transform of the amplitude spectrum provides the group delay (Fig. 22(d)). The similarity between Figs. 22(c) and Fig. 22(d) is excellent. This is the reason why the phase spectrum of the pitch waveform need not be encoded.

We anticipated that the new 2400-b/s would rank somewhere between the 2400-b/s LPC and the 4800-b/s CELP. This turned out to be true with female speech with no noise, as illustrated in Fig. 23. Our new 2400-b/s voice encoder was more intelligible than the existing 2400-b/s. We will continue to refine our new speech encoder, and further test scores will be published.



(a) Female speech in quiet environment

(b) Female speech in noisy environment (with destroyer noise)

Fig. 23 — DRT scores of our new speech encoder are compared to those of the existing 2400-b/s LPC and 4800-b/s CELP. As noted, our new 2400-b/s voice encoder is better than the existing 2400-b/s LPC.

Table 7 — Comparison of New Speech Analysis/Synthesis Technique with Existing Techniques

	Existing Techniques	New Technique
Speech Waveform	The speech waveform is treated as a continuous time sequence. Analysis often begins with a regression analysis, correlation analysis, or filtering.	The speech waveform is considered as a collection of disjoint waveforms. Analysis begins with pitch-synchronous segmentation of individual pitch waveforms.
Speech Modeling	The speech production mechanism (vocal tract, glottis, etc.) is modeled.	The speech waveform itself is modeled.
Pitch Interference	Pitch interference (periodic discontinuities of the speech waveform) adversely affects speech spectral envelope estimation.	Pitch interference has no effect on the analysis because the individual pitch waveforms have been segmented prior to analysis.
Speech Alterations	Only synthetic speech (vocoded speech) can be altered.	The original speech waveform can be altered directly.

CONCLUSIONS

This report describes a new speech analysis/synthesis technique that is a major departure from existing speech analysis/synthesis techniques. The differences between the new and existing techniques are summarized in Table 7.

One important application of our new speech analysis/synthesis technique is in tactical message systems, which transmit speech in terms of words, phrases, and text. These speech entities are denoted by a symbol; speech is regenerated at the receiver by concatenating the raw speech corresponding to each received symbol. The pitch contour or speech rate can be altered by our new speech analysis/synthesis technique to produce more natural-sounding speech Because the required data rate is well below 75 b/s, such a voice message system is well-suited for burst or stealth voice communication. It is also practical for implementing a secure voice underwater communication system operating between submarines, or between submarines and surface ships.

Another important application of our new speech analysis/synthesis technique is in the implementation of a new low-data-rate vocoder operating at 2400 b/s or below. According to our initial tests, encoded speech at 2400 b/s is better for female speech in a quiet environment than the existing 2400-b/s LPC. This is also true in noisy environments. Because Navy tactical platforms are generally noisy and female members are increasingly being deployed in the battlefield, our 2400-b/s speech encoder should be able to achieve more effective communication than the existing 2400-b/s LPC.

ACKNOWLEDGMENTS

We thank the Research Advisory Committee of the Naval Research Laboratory for sponsoring this research effort. We also thank I. Smietan and T. McChesney of SPAWAR PD71 for their support of this and other related R&D efforts. Without this support, we could not have pursued this research. Finally, the authors also express thanks to Stephanie Everett of NRL who read the manuscript and provided invaluable comments.

REFERENCES

1. Joint Staff, "C4I for the Warrior," Pentagon, Washington, DC 20318-6000 (1992).

2. Joint Staff, "Top Five Future Joint Warfighting Capabilities," unidentified briefing viewgraph, Pentagon, Washington, DC 20318-6000 (1994).

3. Chief of Naval Research, "ONR Naval Needs and Scientific Opportunities," Office of Naval Research, Memo Ser 10P2/1707 (1992)..

4. Naval Space and Electronic Warfare Office, "Implementation of Copernicus Architecture Requirements Definition, Appendix A-5," Chief of Naval Operations, Memo Ser 941C/ IU552769 (1991).

5. A. Schmidt-Nielsen, "Identifying Familiar Talkers Over a 2.4 kbps LPC Voice System," *J. Acoust. Soc. Am.*, **75**, S60 (1984).

6. H. Dudley, "The Carrier Nature of Speech," *Bell System Tech. J.* **19**, 495-515 (1940).

7. T. Tremain, "Government-Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology* **1(2)**, 40-44 (1982).

8. A. Papoulis, *Signal Analysis* (McGraw-Hill Book Company, New York, 1977).

9. G.S. Kang, T.M. Moran, and D.A. Heide, "Voice Message Systems for Tactical Applications (Canned Speech Approach)," NRL Report 9569 (1993).

10. Prof. J. Martin of University of Maryland, personal communication, 1993.

11. R.K. Potter, G.A. Kopp, and H.G. Kopp, *Visible Speech* (Dover Publications, Inc., New York, 1966).

12. A.V. Oppenheim and R.W. Shafer, *Discrete-Time Signal Processing* (Prentice Hall, Englewood Cliffs, NJ, 1989).

13. S. Everett, "Automatic Speaker Recognition Using Vocoded Speech," ICASSP, 383-386 (1985).

14. G.S. Kang and L.J. Fransen, "Second Report of the Multirate Processor (MRP) for Digital Voice Communication," NRL Report 8614 (1982).

15. T. Tremain, "DoD Voice Processor Consortium Report on Performance of the LPC-10e Voice Processor," Government test report (1989).

16. J.S. Carofolo, "ARPA TIMIT Acoustic-Phonetic Speech Database," National Institute of Standards and Technology, Gaithersburg, MD 20899.

Appendix

2400-b/s LPC with Pitch-Synchronous Residual Excitation

This·appendix describes an alternative method for encoding high-quality speech at 2400 b/s. We include this encoder to show that synchronous speech waveform segmentation can be exploited in the existing LPC speech analysis/synthesis system. The two speech encoders differ by:

- The 2400-b/s speech encoder described in the main part of this report quantizes directly the pitch-synchronously segmented speech waveform [Fig. A1(a)]. This is a direct method; a portion of the speech waveform is directly quantized.

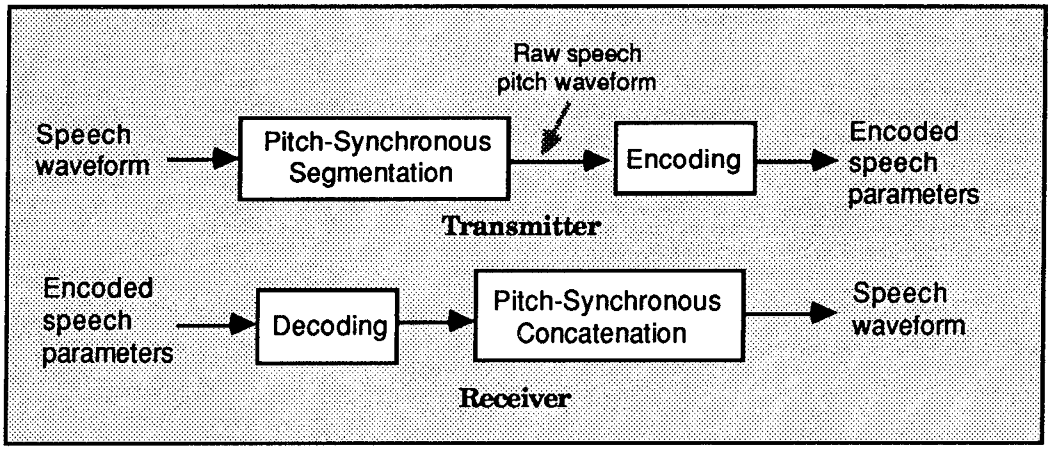
- The 2400-b/s speech encoder presented in this Appendix quantizes the pitch-synchronously segmented prediction residual after LPC analysis is performed on the speech waveform [Fig. A1(b)]. This alternative approach is an extension of the residual-excited LPC.

Residual-Excited LPC

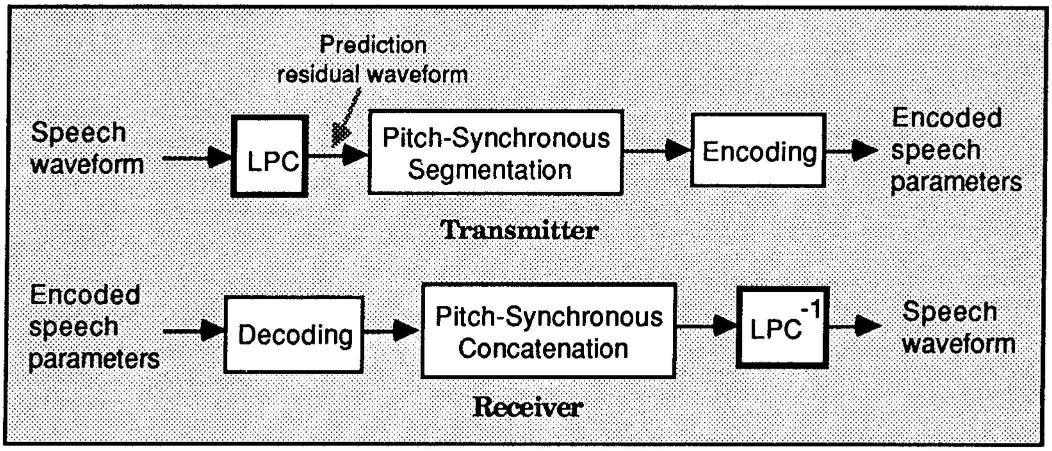
A residual-excited LPC is an LPC speech/synthesis system where the excitation signal is the prediction residual computed from Eq. (20). In the frequency domain, the residual spectrum is a product of the speech spectrum and the LPC filter response. Thus, the prediction residual is an ideal excitation signal for the LPC analysis/synthesis system. In the absence of parameter quantization, the prediction residual produces an output identical to the input.

It is significant to note that the deficiencies of LPC analysis (discussed earlier) will not be reflected in the output speech of a residual-excited LPC because the prediction residual contains the inverse of those deficiencies to remove their effects. Such is not true with the pitch-excited LPC, where the excitation signal is not coupled with the input speech. On the other hand, the pitch-excited LPC is capable of altering speech characteristics (i.e., speech rate, pitch period, or resonant frequencies), similar to other speech models. The residual-excited LPC, however, is not a speech model in the strict sense, and the speech characteristics cannot be altered.

The residual-excited LPC, however, is capable of mimicking the input speech (i.e., vocoding). The output speech quality is higher than the pitch-excited LPC [A1], but the data rate is also higher. Note that even if only one bit is allocated for each residual sample, the data rate to encode the prediction residual alone would be 8 kb/s for the speech sampling rate of 8 kHz. Speech generated by the prediction residual with one per sample is not that good. Thus, various methods have been developed to minimize the number of bits to encode the prediction residual. One such method encodes the prediction residual for low-frequency components only (typically 0 to 1 kHz) and regenerates high frequencies at the receiver. This approach is feasible because the prediction residual has a relatively flat spectral envelope (Fig. A2). In this way, each low-frequency prediction residual can be quantized with up to three bits.


Raw speech pitch waveform
Speech waveform
Pitch-Synchronous Segmentation
Encoding
Encoded speech parameters
Transmitter
Encoded speech parameters
Decoding
Pitch-Synchronous Concatenation
Speech waveform
Receiver

(a) Direct approach (discussed in the main part of this report)


Prediction residual waveform
Speech waveform
LPC
Pitch-Synchronous Segmentation
Encoding
Encoded speech parameters
Transmitter
Encoded speech parameters
Decoding
Pitch-Synchronous Concatenation
LPC^{-1}
Speech waveform
Receiver

(b) Residual-excited LPC approach (discussed in this Appendix)

Fig. A1 — High-level block diagrams indicate differences between the two 2400-b/s voice processors discussed in this report. Figure A1(a) is a direct method where the pitch-synchronously segmented speech waveform is directly encoded. Figure A1(b) is a modified residual-excited LPC where the pitch-synchronously segmented prediction residual is encoded.

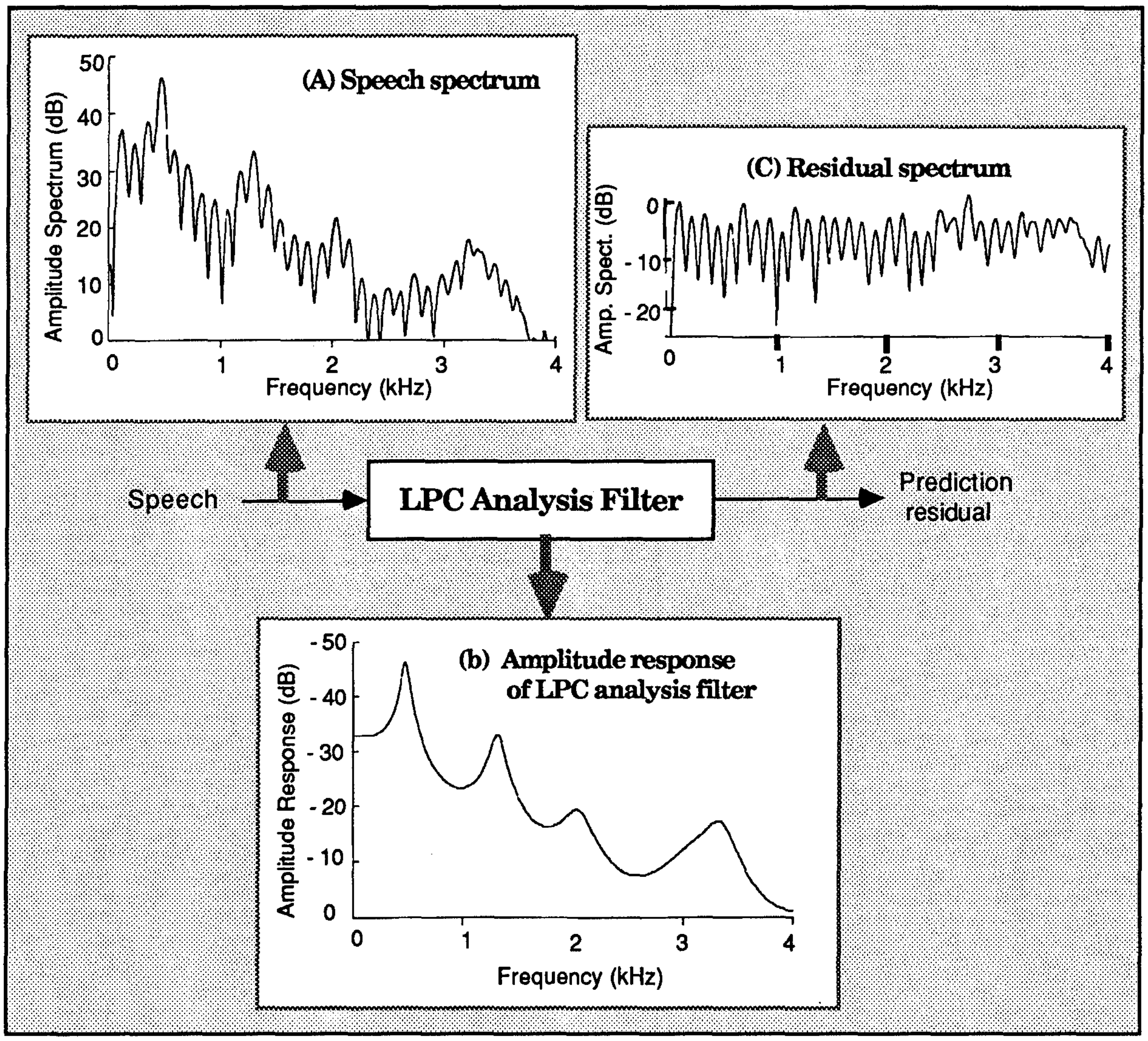


Fig. A2 — Speech spectrum, residual spectrum, and LPC analysis filter response. As noted, the residual spectrum is a complementary part of the speech spectrum unaccounted by the LPC analysis filter response.

Because our target data rate is as low as 2400 b/s, we require more effective bit-saving measures than have been available in the past. The two techniques used in our 2400-b/s residual-excited LPC are:

- *Pitch-synchronously segmented residual*: The prediction residual is pitch-synchronously segmented. The amplitude spectrum of the segmented waveform is then encoded by vector quantization. The phase spectrum is regenerated from the amplitude information.

- *Vector quantization of line-spectrum pairs (LSPs)*: The LPC coefficients are converted to LSPs. The resultant LSPs are quantized semilogarithmically at 12 steps per octave. All 10 LSP frequencies are encoded by two look-up tables.

Figure A3 is a block diagram of the residual-excited LPC with the above-mentioned features.

Bit Allocations

Frame size is 180 samples (or 22.5 ms). Therefore, 54 bits are available from each frame to encode speech information. The synchronization bit is alternating "1"s and "0"s. From previous experience, we have found five bits to be adequate for quantizing the pitch period or speech rms value (Table A1). We separately tested for the number of bits required to encode LSPs in two tables and found 28 bits to be sufficient. Remaining bits are used to encode the prediction residual samples.

Pitch Period Encoding

The pitch period is semilogarithmically quantized into 20 steps per octave from a pitch period of 20 to 156 (i.e., pitch frequencies from 400 Hz to 51 Hz). The total number of pitch steps is 60 which requires five bits for encoding. Table A2 is the pitch encoding/decoding table.

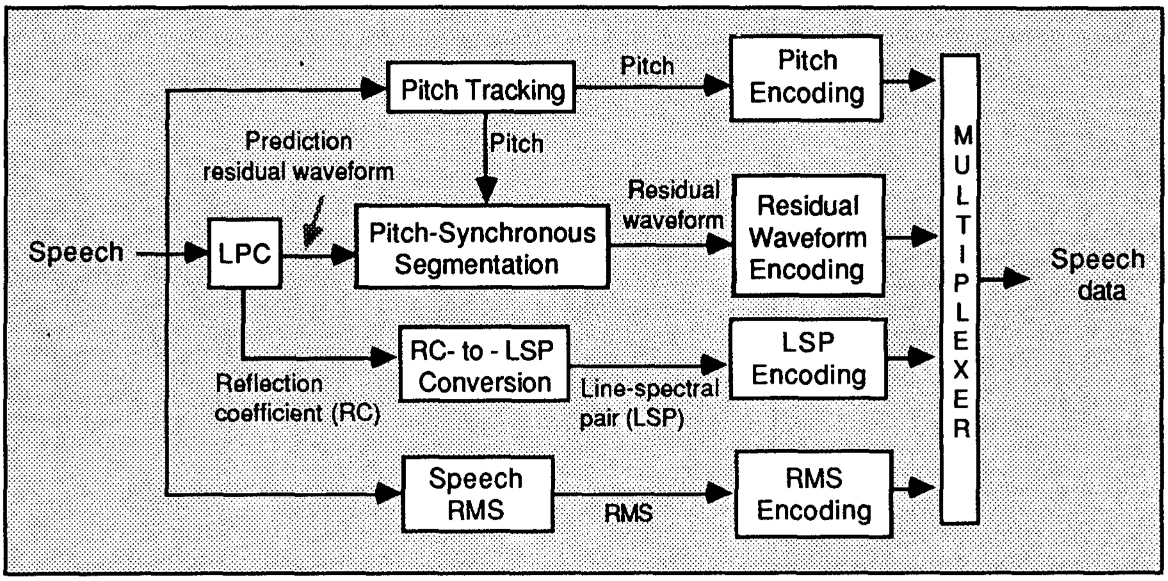
Speech Rms Encoding

Speech rms values ranging from 0 to 511 (9-bit quantity) is compressed to a 5-bit quantity through a logarithmic transformation. Table A3 is the rms coding/decoding table. This table is identical to that used in the current 2400-b/s LPC.

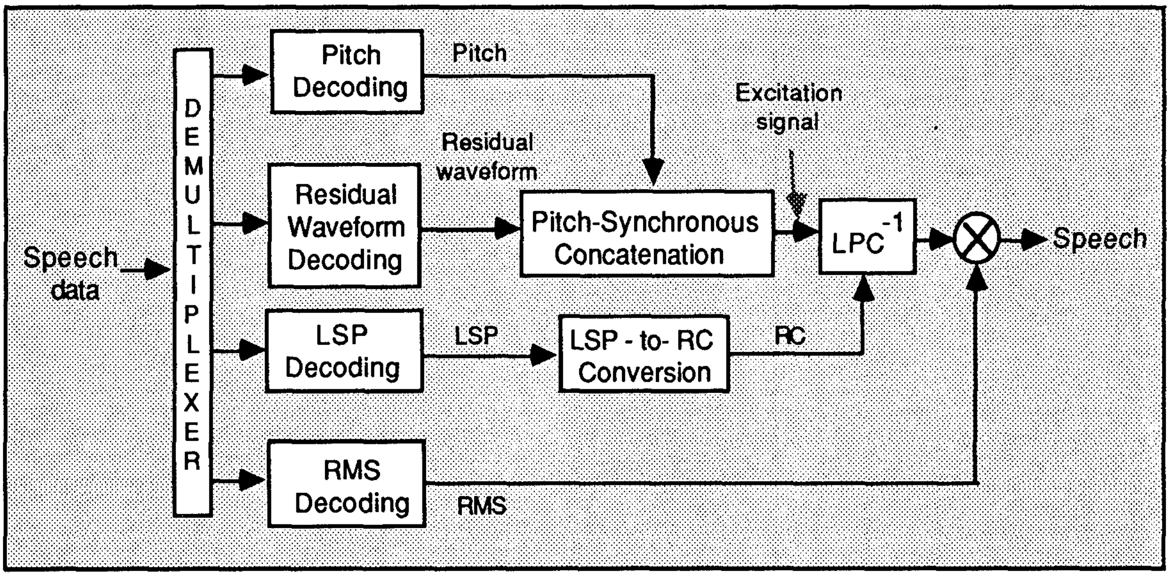
LSP Encoding

LSPs are obtained by transforming LPC coefficients (prediction coefficients or reflection coefficients) generated by the linear prediction analysis. References A2 and A3 describe the procedures for converting from LPC coefficients to LSPs. These reports also discussed the advantages of using LSPs over prediction coefficients or reflection coefficients. LSPs are now widely used as speech parameters in various low-data-rate speech encoders, such as the 4800-b/s codebook-excited linear predictor (CELP).

As a first step to LSP coding, each LSP frequency is represented by one of the 40 logarithmically quantized frequencies at 12 steps per octave (Table A4). According to our tests, speech quality is virtually unimpaired by quantizing frequency-domain speech parameters (e.g., pitch frequency, formant frequency, etc.) into 12 steps per octave.



(a) Transmitter



(b) Receiver

Fig. A3 — Block diagram of a 2400-b/s LPC with pitch-synchronously-segmented residual excitation. If the excitation shown in this figure is replaced by the conventional pitch-excitation signal, the resulting voice encoder is very similar to the existing 2400-b/s LPC, but the data rate would be around 1800 b/s.

Table A1 — Bit Allocation for Each Frame

	(bits)
Synchronization	1
Pitch period	5
Speech rms	5
Filter coefficients	28
Pitch-synchronously segmented residual waveform	15
Total	54

Table A2 — Pitch Encoding/Decoding Table

Pitch Period		
Encoded Value	Code	Decoded Value
20 or less	0	20
21	1	21
22	2	22
23	3	23
24	4	24
25	5	25
26	6	26
27	7	27
28	8	28
29	9	29
30	10	30
31	11	31
32	12	32
33	13	33
34	14	34
35	15	35
36	16	36
37	17	37
38	18	38
39	19	39
40-41	20	40
42-43	21	42
44-45	22	44
46-47	23	46
48-49	24	48
50-51	25	50
52-53	26	52
54-55	27	54
56-57	28	56
58-59	29	58
60-61	30	60
62-63	31	62
64-65	32	64
66-67	33	66
68-69	34	68
70-71	35	70
72-73	36	72
74-75	37	74
76-77	38	76
78-79	39	78
80-83	40	81
84-87	41	85
88-91	42	89
92-95	43	91
96-99	44	97
100-103	45	101
104-107	46	105
108-111	47	109
112-115	48	113
116-119	49	117
120-123	50	121
124-127	51	125
128-131	52	129
132-135	53	133
136-139	54	137
140-143	55	141
144-147	56	146
148-151	57	149
152-155	58	120
156 or more	59	120

Table A3 — Speech Rms Coding/Decoding Table

Speech Amplitude		
Encoded Value	Code	Decoded Value
0	0	0
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9-10	9	9
11-12	10	11
13-14	11	13
15-17	12	16
18-21	13	19
22-25	14	23
26-30	15	27
31-35	16	32
36-42	17	39
43-51	18	46
52-60	19	55
61-72	20	66
73-86	21	79
87-103	22	94
104-123	23	113
124-147	24	135
148-176	25	164
177-210	26	192
211-251	27	230
252-300	28	275
301-359	29	328
360-428	30	392
429-511	31	468

Table A4 — Representation of Individual LSP Frequencies*

LSP (Hz)		
Encoded Value	Code	Decoded Value
400-423	0	411
424-448	1	435
449-475	2	461
476-503	3	489
504-533	4	518
534-565	5	549
566-598	6	581
599-634	7	616
635-672	8	653
673-712	9	692
713-754	10	733
755-799	11	776
800-847	12	776
848-897	13	823
898-950	14	872
951-1007	15	923
1008-1067	16	978
1068-1130	17	1037
1131-1198	18	1098
1199-1269	19	1164
1270-1344	20	1233
1345-1424	21	1306
1425-1509	22	1383
1510-1599	23	1466
1600-1694	24	1646
1695-1795	25	1744
1796-1902	26	1848
1903-2015	27	1958
2016-2135	28	2074
2136-2262	29	2198
2263-2396	30	2328
2397-2539	31	2466
2540-2690	32	2613
2691-2850	33	2769
2851-3019	34	2933
3020-3199	35	3108
3200-3389	36	3293
3390-3591	37	3489
3592-3804	38	3696
3805-4000	39	3900

* Note: This LSP quantization table is for the residual-excited LPC where a lack of low frequencies (below 400 Hz) is compensated by the prediction residual. For a pitch-excited LPC, the table should contain several frequencies below 400 Hz.

The most efficient way of encoding LSPs is to represent all 10 LSP frequencies vectorially by a single code. This approach, however, is impractical because the table size becomes as large as 847,660,528, which is the total number of combinations for choosing 10 out of 40 frequencies. Therefore, we use two tables; the first table lists all possible combinations of the first four LSPs, and the second table lists the following six LSPs. The number entries in either table is $2^{14} = 16{,}384$ if we limit the range of each LSP based on an LSP distribution generated from a large speech database (Table 5).

The two LSP look-up tables are constructed by using the information contained in Tables A4 and A5. For example, the first LSP look-up table is a four-dimensional table, $T_1(f_1, f_2, f_3, f_4)$ where f_1, f_2, f_3, and f_4 are the first through fourth LSP frequencies, respectively, listed in Table A5. Likewise, the second LSP look-up table is a six-dimensional table, $T_2(f_5, f_6, f_7, f_8, f_9, f_{10})$. Either table has 16,384 (= 2^{14}) frequency sets. Thus, the number of bits required to encode 10 LSP frequencies is 28 bits.

Table A5 — Range of Each LSP in the Two LSP Tables

	Table 1	Table 2
LSP1	1 to 12	
LSP2	3 to 19	
LSP3	12 to 26	
LSP4	15 to 29	
LSP5		21 to 32
LSP6		25 to 34
LSP7		28 to 36
LSP8		31 to 37
LSP9		34 to 39
LSP10		36 to 40

Residual Encoding

The most important part of the 2400-b/s encoder under discussion is in the residual encoding. Because we have only 667 b/s to encode the prediction residual, this 2400-b/s voice encoder cannot achieve the performance level of the high-data-rate residual-excited LPC, but a careful design of residual encoding can accomplish higher performance over existing 2400-b/s LPCs that do not encode the residual information.

In this new residual-excited LPC, the prediction residual is generated for each frame from the speech waveform and the quantized LSPs. In the absence of residual quantization, the output still equals the input, even if LSPs are quantized. The prediction residual is then pitch-synchronously segmented by the technique presented in the main part of this report. The segmented prediction residual for one pitch period is encoded and transmitted. The segmented prediction residual may be encoded in the time domain or frequency domain. We chose the frequency-domain encoding because of the following reasons:

- *Insignificance of pitch-synchronously segmented residual phase spectrum*: In the previous residual-excited LPC, residual samples are time-referenced at the analysis frame epoch. Therefore, the quantized phase spectrum critically affects the pitch-to-pitch regularity of the time waveform. If the phase spectrum is coarsely quantized, synthesized speech invariably becomes wobbly as the result of pitch-to-pitch waveform jitters. In our approach, the residual waveform is time-referenced at the pitch epoch, and the residual waveform is segmented for one pitch period. Therefore, the phase spectrum is not as critical. We can substitute the actual phase spectrum with artificial phase spectrum; this is discussed later.

- *Amplitude spectrum lacks pitch harmonics*: The amplitude spectrum of the residual from the previous residual-excited LPC has a relatively flat spectral envelope, but pitch harmonics are inscribed under the spectral envelope. Often a notch filter is used to suppress pitch harmonics. A coarse quantization of the amplitude spectrum results in raspy output speech. In contrast, the amplitude spectrum of pitch-synchronously segmented residual (Fig. A4) is free from pitch harmonics, and coarse quantization produces more tolerable synthesized speech.

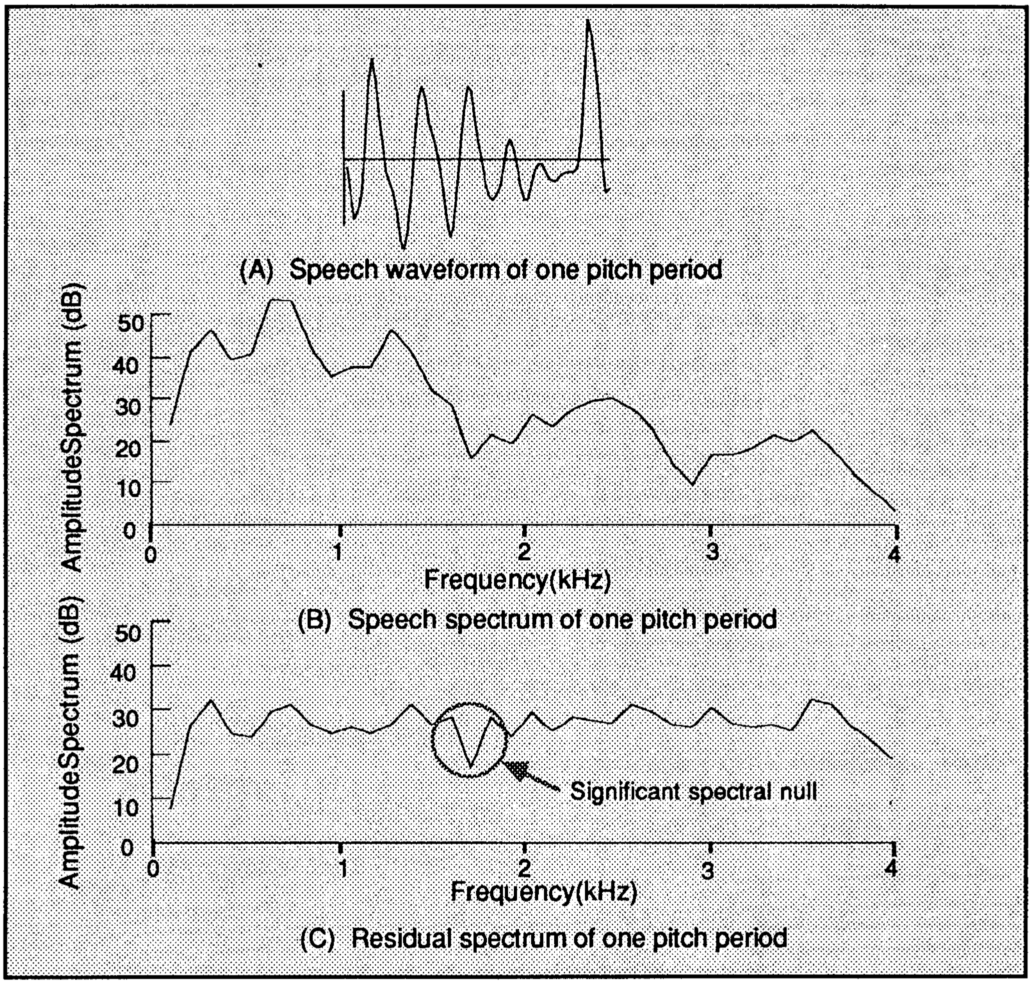


Fig. A4 — Speech waveform, speech spectrum, and residual spectrum for one pitch period. As noted, either the speech or residual amplitude spectrum lacks pitch harmonics. A spectral dip (spectral zero) unfiltered by the LPC analysis filter is significant. It should be captured in the residual encoding to generate higher quality speech.

- *Frequency-selective quantization is possible*: Human ears are more tolerant to quantization errors in higher frequencies. By encoding the residual in the frequency domain, we can exploit this human perception characteristic.

Amplitude Spectrum Representation

Fifteen bits per frame are available to encode the amplitude spectrum. Clearly, we cannot encode too many spectral components with only 15 bits. After several test runs, we decided to encode the first 12 amplitude spectral components. For high-pitch female voices, 12 amplitude spectral components covers nearly the entire passband. On the other hand, for low-pitch male voices, these 12 amplitude spectral components cover only the lower frequency band. During synthesis, missing amplitude spectral components are replaced by the average value of the spectral components transmitted. This is permissible because the pitch-synchronously segmented residual has a relatively flat amplitude spectrum without being modulated by pitch harmonics. The 12 amplitude spectral components are subjected to vector quantization. The procedure is identical to what has already been discussed in *Pitch Waveform Quantization* in the main part of this report.

Phase Spectrum Representation

As stated earlier, we do not encode the phase spectrum. The actual phase spectrum is replaced by an artificial phase spectrum based on the speech waveform periodicity. We note that all voiced speech waveforms have the first resonant frequency (denoted by F1). If the first resonant frequency is present, the phase spectrum is time-invariant; otherwise, the phase spectrum is random from frame to frame.

The presence or absence of the first resonant frequency is reliably detected by the the location of the first two LSP frequencies (Fig. A5). As noted, when speech is voiced, these two frequencies are located close to each other, somewhere below 800 Hz. We use the product of the sum and difference of these two frequencies to indicate the presence or absence of the first resonant frequency (Fig. A5).

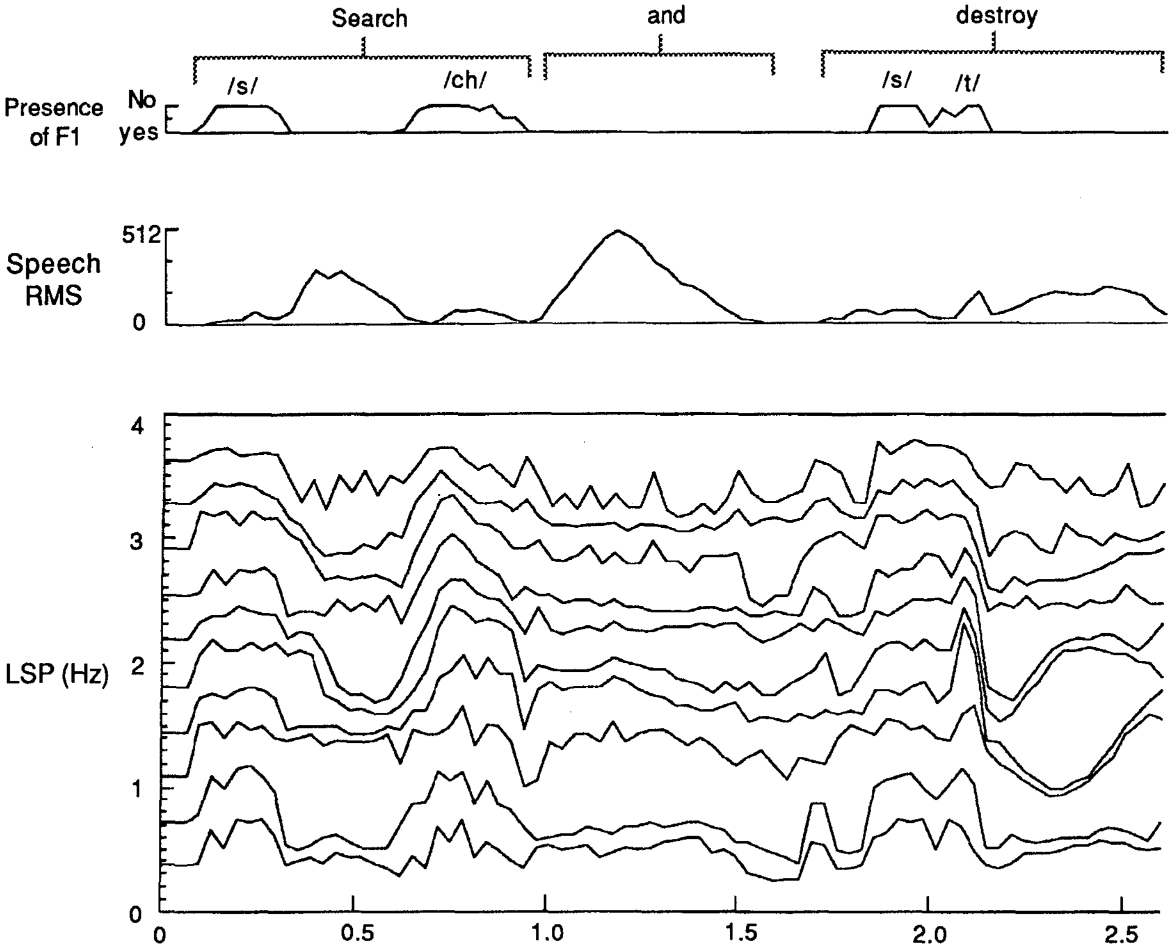


Fig. A5 — LSP trajectories of "Search and Destroy." As noted when the speech is voiced, the first two LSP frequencies are located closely somewhere below 800 Hz. A product of sum and difference of these two frequencies reliably indicates the presence or absence of the first resonant frequency (see the uppermost plot, designated as "Presence of F1").

If the first resonant frequency is present in speech, the phase spectrum is stationary (i.e., time invariant), and it is a quadratic function of frequency to achieve low peak amplitudes [A4]:

$$\phi_o(\kappa) = (2\pi)\xi(\kappa/K)^2 \qquad \kappa = 1, 2, ..., K, \tag{A1}$$

where $\phi_o(k)$ is the kth stationary phase component, κ is a frequency index, and K is the total number of spectral components related to the pitch period. See Eq. (12) of the main part of this report for K when the pitch period is an odd number. The quantity ξ is an integer number, larger ξ corresponds to smaller peak amplitudes. We use $\xi = 2$.

On the other hand, if the first resonant frequency is absent, a random phase spectrum expressed by

$$\Delta\phi(\kappa) = (\pi/2)\ \sigma(k)\ (\kappa/K) \qquad \kappa = 1, 2, ..., K, \tag{A2}$$

where $\sigma(k)$ is a uniformly distributed random variable between –1 and 1, κ is a frequency index, and K is the total number of spectral components within the 0-4 kHz passband.

Note that the F1 indicator is not the voicing decision used in the pitch-excited vocoder, although they may be similar in certain cases. For example, if the speech is a fricative, the first resonant frequency will be absent. In this case, we use the phase spectrum defined by Eq. (A2), which is similar to that for an unvoiced excitation signal used in the pitch-excited vocoder. For many other consonants (e.g., /p/, /t/, etc.), however, the first resonant frequency will be present, and we use the phase spectrum defined by Eq. (A1), rather than by Eq. (2). For these consonants, the residual amplitude spectrum plays a critical role.

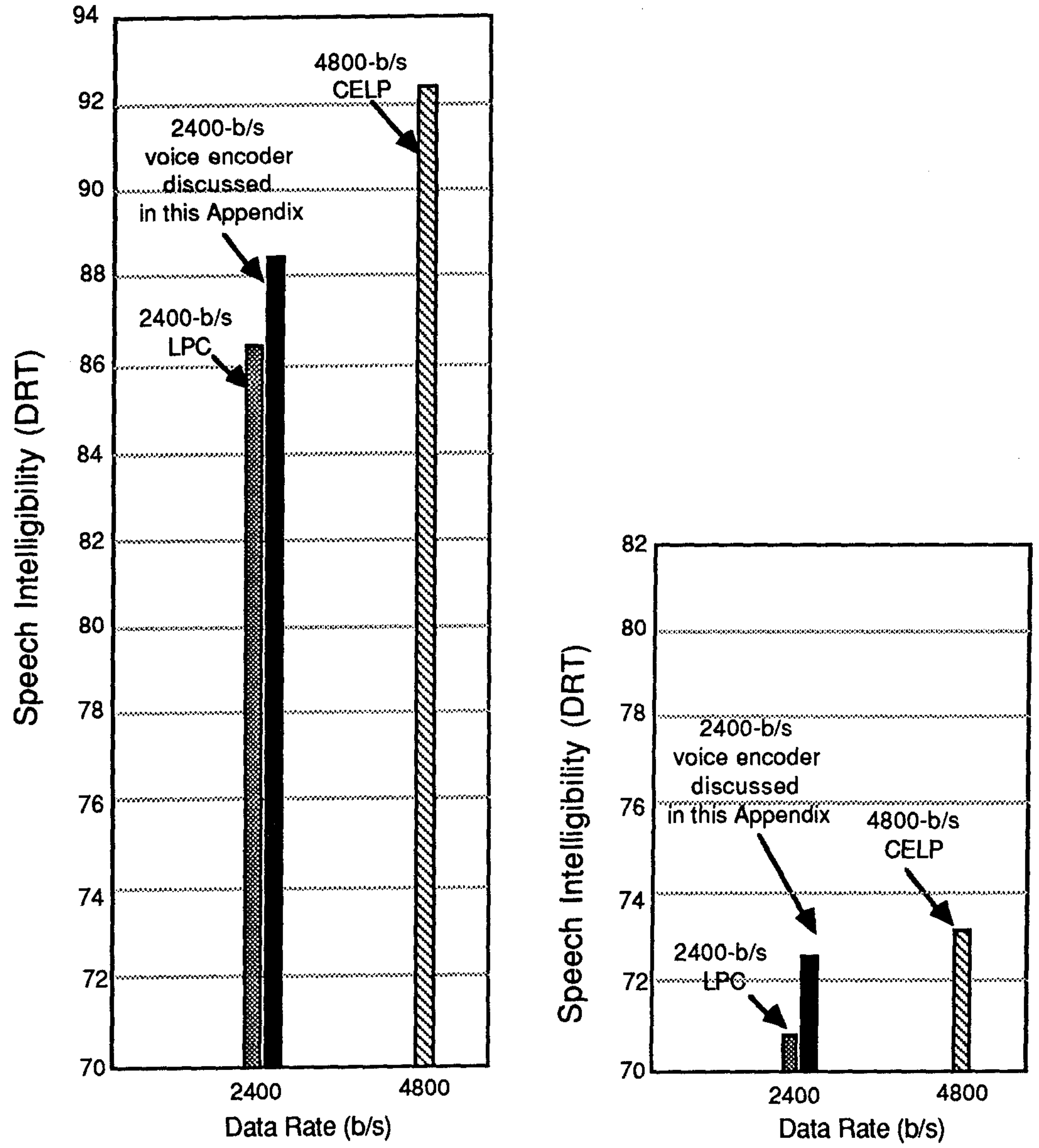
Intelligibility Scores

Any new 2400-b/s speech encoder must outperform the existing 2400-b/s LPC, especially in difficult operating conditions such as female speech and noisy speech. We processed DRTs for these two cases. The voice processors we compared were: 2400-b/s LPC; our new 2400-b/s voice encoder; and 4800-b/s CELP. We anticipated that the new 2400-b/s would rank somewhere between the 2400-b/s LPC and 4800-b/s CELP. This turned out to be true with female speech with no noise, as illustrated in Fig. A6.

Summary

The pitch-synchronous speech segmentation method presented in the main part of this report can be exploited in the residual-excited LPC to improve speech intelligibility. In this application, the prediction residual is pitch-synchronously segmented, and the resultant waveform is encoded. At the receiver, the segmented residual waveform is concatenated to become the excitation signal.

The speech intelligibility of this voice encoder operating at 2400 b/s is approximately two points better than the current 2400-b/s LPC. A DRT improvement of two points is significant. It could upgrade the speech intelligibility from "good" to "very good" or from "moderate" to "good," according to the intelligibility classification used by the DoD Voice Processor Consortium.



(a) Female speech in quiet environment

(b) Female speech in noisy environment (with destroyer noise)

Fig. A6 — DRT scores of our new speech encoder are compared to those of the existing 2400-b/s LPC and 4800-b/s CELP. As noted, our new 2400-b/s voice encoder is better than the existing 2400-b/s LPC.

References

A1. G.S. Kang and L.J. Fransen, "Second Report of the Multirate Processor (MRP) for Digital Voice Communications," NRL Report 8614 (1982).

A2. G.S. Kang and L.J. Fransen," Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)," NRL Report 8857 (1985).

A3. G.S. Kang and L.J. Fransen, "High-Quality 800-b/s Voice Processing Algorithm," NRL Report 9301 (1991).

A4. G.S. Kang and S. Everett, "Improvement of the Excitation Source in the Narrowband Linear Prediction Vocoder," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **ASSP-33**(2), 377-386 (1985).